Dr. Mathias Jesussek

Dr. Hannah Volk-Jesussek

Statistics made easy

A clear and simple introduction

6th edition

©DATAtab e.U. | Graz | 2025

The work including all its parts is protected by copyright. Any use not expressly permitted by copyright law requires the prior consent of DATAtab e.U.

Cover design: Mathias Jesussek

Introduction 19
1. Descriptive statistics and inferential statistics 20
1.1 Subareas of statistics 20
1.2 Descriptive statistics 21
1.3 Inferential Statistics 25
2. Level of measurment 29
2.1 Nominal variables 29
2.2 Ordinal variables 30
2.3 Categorical variables 31
2.4 Metric variables 31
2.5 Ratio scale and interval scale 32
3. Sampling 33
3.1 Full or total survey vs. random sample 33
3.2 Population and sample 34
3.3 Types of sampling? 34
3.4 Probability selection 35
3.5 Deliberate selection 35
3.6 Arbitrary selection 36
3.7 Sample selection in online surveys 36
3.8 Sample description in bachelor or master thesis 36
4. Location parameter 38
4.1 Mean value (arithmetic mean) 39
4.2 Geometric mean and quadratic mean 40
4.3 Median 40
4.4 Mean and median in comparison 42
4.5 Mode (Modal value) 42
4.6 Advantage and disadvantage of the Mean, Median and Mode 44
5. Dispersion parameter 47
5.1 Standard deviation 48
5.2 Variance 49
5.3 Difference between variance and standard deviation 50
5.4 Range 51
5.5 Quartile 52
5.6 Interquartile range 53
5.7 Example dispersion parameter 53
6. Frequency table 56

6.1 Absolute and relative frequencies ... 56
6.2 Valid percent ... 57
6.3 Frequency table in statistics ... 58
6.4 Example frequency table ... 59
6.5 Frequency table in APA style ... 61
7. Contingency table (Crosstab) ... 62
7.1 Crosstabs in statistics ... 62
7.2 Interpretatation of crosstabs ... 63
7.3 Example crosstab ... 64
7.4 Testing a crosstab for significance ... 65
8. Charts ... 66
8.1 Bar chart ... 67
8.2 Bar chart for frequencies ... 68
8.3 Grouped bar charts ... 69
8.4 Bar chart for mean values ... 70
8.5 Error bar ... 70
8.6 Example bar chart ... 71
8.7 Histogram ... 73
8.8 Histogram example ... 74
8.9 Bar chart vs. Histogram ... 75
8.10 Scatter plot ... 76
8.11 Line charts ... 77
8.12 Boxplot ... 78
8.13 Bland-Altman plot ... 84
8.13.1 Example of Bland-Altman plot ... 85
8.13.2 Structure of a Bland-Altman plot ... 85
8.13.3 How can a Bland-Altman plot be used? ... 86
8.13.4 Find outliers in the data ... 86
8.14 Create charts online with DATAtab ... 87
9. Inferential Statistics ... 90
9.1 Hypotheses ... 90
9.2 Null and alternative hypothesis ... 93
9.3 Difference and correlation hypotheses ... 94
9.4 Directional and undirectional hypotheses ... 96
9.5 Hypothesis Testing ... 99
9.5.1 Hypothesis testing and the null hypothesis ... 100

9.5.2 The uncertainty in hypothesis testing 101
9.5.3 Level of significance or probability of error 101
9.5.4 Example significance level and p-value 103
9.5.5 Types of errors 103
9.5.6 Significance vs effect size 104
9.5.7 Choosing the appropriate hypothesis test 105
9.5.8 Examples for hypothesis tests 108
9.6 The p-value 109
9.6.1 Defining the p-value 109
9.6.2 Using the p-value 110
9.6.3 Significance level 111
9.6.4 One-tailed p-values 112
9.6.5 Calculate p-value 113
9.6.6 Statistical tests and the p-value 114
9.6.7 Specify the p-value 115
10. Checking assumptions of statistical tests 117
10.1 Levene test of variance homogeneity 117
10.2 Levene test example 119
10.3 Interpreting the Levene Test 121
10.4 Normality test 122
10.4.1 Statistical test for normal distribution 122
10.4.2 Disadvantage of analytical tests for normal distribution 124
10.4.3 Graphical test for normal distribution 125
10.5 Multicollinearity test 127
10.5.1 How to avoid multicollinearity? 128
10.5.2 Multicollinearity test 129
10.5.3 Tolerance value 129
10.5.4 VIF Multicollinearity 130
11. Statistical tests for differences 131
11.1 One sample t-test 131
11.1.1 Basics of the one sample t-test 131
11.1.2 Examples of a t-test for one sample 132
11.1.3 Assumptions of the one-sample t-test 133
11.1.4 Hypotheses for the one-sample t-test 134
11.1.5 Calculation of the one-sample t-test 135
11.1.6 One sample t-test with example 137

11.1.7 APA format | One-sample t-test ... 140
11.2 T-test for independent samples (unpaired t-test) ... 141
11.2.1 Using an independent t-test ... 141
11.2.2 Purpose of the independent/unpaired t-test ... 142
11.2.3 Examples for the unpaired t-test ... 143
11.2.4 Research question and hypotheses for the unpaired t-test ... 143
11.2.5 Assumptions unpaired/independent t-test ... 145
11.2.6 Calculate t-test for independent samples ... 146
11.2.7 Confidence interval for the true mean difference ... 148
11.2.8 One-sided and two-sided unpaired t-test ... 149
11.2.9 Effectsize unpaied t-test ... 149
11.2.10 Example t-test for independent samples ... 150
11.2.11 Interpretation t-test for independent samples ... 153
11.2.12 Report a t-test for independent samples ... 154
11.3 Paired-samples t-test ... 157
11.3.1 Why do you need the paired t-test? ... 157
11.3.2 What is the advantage of a dependent t-test over an independent t-test? ... 159
11.3.3 Examples of the t-test for paired samples ... 159
11.3.4 Research question and hypotheses of the paired t-test ... 160
11.3.5 Assumptions paired t-test ... 161
11.3.6 Calculating a paired t-test ... 162
11.3.7 Example t-test for dependent samples with DATAtab ... 163
11.3.8 Interpretation of a t-test for dependent samples ... 166
11.3.9 Effect size dependent t-test ... 166
11.4 Mann-Whitney U test ... 167
11.4.1 Assumptions Mann-Whitney U test ... 168
11.4.2 Hypotheses Mann-Whitney U test ... 168
11.4.3 Calculate Mann-Whitney U test ... 169
11.4.4 Calculate Mann-Whitney U test with tied ranks ... 171
11.4.5 Mann-Whitney U test Example with DATAtab ... 173
11.4.6 Interpret Mann-Whitney U test ... 177
11.4.7 Mann-Whitney U test and effect size ... 177
11.5 Wilcoxon test ... 178
11.5.1 Assumptions of the Wilcoxon test ... 179
11.5.2 Hypotheses in the Wilcoxon test ... 179
11.5.3 Wilcoxon test and test power ... 180

11.5.4 Calculate Wilcoxon test ... 180
11.5.5 Calculate Wilcoxon signed-rank test with tied ranks ... 185
11.5.6 Effect size in the Wilcoxon signed-rank test ... 187
11.5.7 Example Wilcoxon test with DATAtab ... 188
12. Frequency analysis ... 190
12.1 Binomial test ... 190
12.1.1 Hypotheses in binomial test ... 190
12.1.2 Binomial test calculation ... 191
12.1.3 Binomial test example ... 191
12.1.4 Interpretation of a Binomial Test ... 193
12.2 Chi-square test ... 194
12.2.1 Applications of the Chi-Square Test ... 195
12.2.2 Calculation of Chi-Square- test ... 196
12.2.3 Chi-Square Test of Independence ... 197
12.2.4 Chi-square distribution test ... 199
12.2.5 Chi-square homogeneity test ... 200
12.2.6 Effect size for Chi-square test ... 200
12.2.7 Effect size vs. p-value ... 201
12.2.8 Example: Chi-square test with DATAtab ... 202
13. Statistical tests to test for differences in more than two groups. ... 207
13.1 Analysis of Variance (ANOVA) ... 207
13.1.1 Why not calculate multiple t-tests? ... 208
13.1.2 Difference between one-way and two-way ANOVA ... 208
13.1.3 Analysis of variance with and without repeated measures ... 209
13.2 One-factor ANOVA ... 210
13.3 One-factor ANOVA example ... 210
13.4 Analysis of variance hypotheses ... 212
13.5 Assumptions for one-way analysis of variance ... 213
13.6 Welch's ANOVA ... 214
13.7 Effect size Eta squared (η^2) ... 214
13.8 Two factor analysis of variance ... 214
13.9 Calculate example with DATAtab ... 215
13.10 Repeated Measures ANOVA ... 217
13.10.1 What are dependent samples? ... 217
13.10.2 Difference of analysis of variance with and without repeated measurements ... 218
13.10.3 Example of repeated measures ANOVA ... 218

13.10.4 Research question and hypotheses ... 219
13.10.5 Assumptions ANOVA with repeated measures ... 220
13.10.6 Results of the one-factor analysis of variance with repeated measures. ... 222
13.10.7 Effect size for repeated measures ANOVA ... 222
13.10.8 Bonferroni Post-hoc-Test ... 223
13.10.9 Calculate ANOVA with measurement repetitions with DATAtab ... 224
13.10.10 Calculate a repeated measures ANOVA by hand ... 225
13.11 Two-way ANOVA (without repeated measures) ... 227
13.11.1 What is a factor? ... 228
13.11.2 Two factors ... 229
13.11.3 Example Two-Way ANOVA ... 230
13.11.4 Hypotheses ... 230
13.11.5 Assumptions ... 231
13.11.6 Calculation of a two-way ANOVA ... 232
13.11.7 Calculating two-way ANOVA with DATAtab ... 236
13.11.8 Interpreting two-way ANOVA ... 239
13.11.9 Interaction effect ... 240
13.12 Two-way ANOVA with measurement repetition ... 242
13.12.1 Sample with measurement repetition ... 243
13.12.2 Example two-way ANOVA with repeated measures ... 244
13.12.3 Hypotheses ... 245
13.12.4 Assumptions of the two-way analysis of variance with repeated measures ... 245
13.12.5 Calculate two-way ANOVA with repeated measures ... 246
13.12.6 Interpret two-way analysis of variance with repeated measures ... 248
13.13 Kruskal-Wallis test ... 250
13.13.1 Examples for the Kruskal-Wallis test ... 250
13.13.2 Research question and hypotheses in the Kruskal-Wallis test ... 251
13.13.3 Assumptions of the Kruskal-Wallis test ... 251
13.13.4 Calculate Kruskal-Wallis test ... 252
13.13.5 Kruskal-Wallis test example ... 254
14. Statistical methods for testing correlations ... 257
14.1 Correlation ... 257
14.1.1 Correlation and causality ... 257
14.1.2 Correlation and causality example ... 258
14.1.3 Correlation interpretation ... 258
14.1.4 Direction of correlation ... 258

14.1.5 Strength of correlation 259
14.1.6 Scatter plot and correlation 260
14.1.7 Test correlation for significance 261
14.1.8 Directional and non-directional hypotheses 262
14.2 Pearson correlation analysis 263
14.2.1 Pearson Correlation assumptions 265
14.3 Spearman rank correlation 266
14.4 Point biserial correlation 267
14.5 Partial correlation 267
14.5.1 Calculation of the partial correlation 268
14.5.2 Partial correlation example 269
14.5.3 Partial correlation 2nd order 270
14.5.4 Example: Pearson correlation 270
14.5.5 Directional (one-sided) correlation hypothesis 274
15. Regression Analysis 275
15.1 Basics of regression 275
15.1.1 Using a regression analysis 276
15.1.2 Types of regression analysis 277
15.1.3 Dummy variables and Reference category 278
15.1.4 Examples of regression: 279
15.2 Linear Regression 281
15.2.1 Simple Linear Regression 282
15.2.2 Multiple linear regression 285
15.2.2.1 Multiple Regression vs. Multivariate Regression 286
15.2.2.2 Coefficient of determination 286
15.2.2.3 Adjusted R2 287
15.2.2.4 Standard estimation error 287
15.2.2.5 Standardized and unstandardized regression coefficient 287
15.2.2.6 Assumptions of Linear Regression 288
15.2.2.7 Linearity 288
15.2.2.8 Homoscedasticity 289
15.2.2.9 Normal distribution of the error 290
15.2.2.10 Multicollinearity 291
15.2.2.11 Significance test and Regression 292
15.2.2.12 Example linear regression 294
15.2.2.13 Interpretation of the results 295

15.2.2.14 Presenting the results of the regression ... 296
15.3 Logistic Regression ... 297
15.3.1 What is logistic regression? ... 298
15.3.2 Logistic regression and probabilities ... 299
15.3.3 Calculate logistic regression ... 299
15.3.4 Logistic function ... 301
15.3.5 Maximum Likelihood Method ... 303
15.3.5.1 The Likelihood Function ... 303
15.3.5.2 Maximum Likelihood Estimator ... 304
15.3.6 Multinomial logistic regression ... 304
15.3.7 Interpretation of the results ... 305
15.3.8 Odds Ratios ... 305
15.3.9 Pseudo-R squared ... 305
15.3.10 Null Model ... 306
15.3.11 Cox and Snell R-square ... 306
15.3.12 Nagelkerkes R-square ... 307
15.3.13 McFadden's R-square ... 307
15.3.14 Chi2 Test and Logistic Regression ... 307
15.3.15 Example logistic regression ... 308
15.3.16 Calculating logistic regression with DATAtab ... 309
15.3.17 Odds Ratios in Logistic Regression ... 311
15.3.18 Basic concept of logistic regression ... 311
15.3.19 Example binary logistic regression ... 311
15.3.20 Odds in logistic regression ... 312
15.3.21 What are the odds? ... 313
15.3.22 What are Odds Ratios? ... 313
15.3.23 Odds Ratio in Logistic Regression ... 315
15.3.24 Calculate Odds Ratios with DATAtab ... 316
15.3.25 Odds ratios of continuous variables ... 318
15.3.26 Odds ratio or exp(B) ... 319
16. Confidence Intervals ... 320
16.1 Why do we need the confidence interval? ... 320
16.2 Interpretation of confidence interval ... 320
16.2.1 Common Misinterpretation of the confidence interval ... 321
16.2.2 Correct interpretation of the CI ... 322
16.3 Calculate confidence interval ... 323

16.4 Confidence interval 95% ... 324
16.5 Confidence interval for t-test ... 325
17. Factor Analysis ... 326
17.1 What is a factor? ... 326
17.2 Example factor analysis ... 327
17.3 Research questions factor analysis ... 328
17.4 Factor load, eigenvalue, communalities ... 329
17.5 Correlation Matrix ... 330
17.6 Factor Analysis and dimensionality ... 330
17.6.1 Eigenvalue criterion (Kaiser criterion) ... 330
17.6.2 Scree-Test ... 331
17.6.3 Communalities ... 332
17.6.4 Component matrix ... 333
17.6.5 Rotation Matrix ... 334
17.6.6 Varimax Rotation ... 334
18. Cluster Analysis ... 335
18.1 Example Hierarchical Cluster Analysis ... 335
18.2 Calculating a Hierarchical Cluster Analysis ... 336
18.3 Distance between two points ... 338
18.3.1 Euclidean Distance ... 338
18.3.2 Manhattan Distance ... 339
18.3.3 Maximum Distance ... 340
18.4 Linking methods ... 341
18.4.1 Single-linkage ... 342
18.4.2 Complete-linkage ... 342
18.4.3 Average-linkage ... 342
18.5 Example Hierarchical Cluster Analysis ... 343
18.5.1 Calculate hierarchical cluster analysis with DATAtab ... 349
18.6 K-means cluster analysis ... 351
18.6.1 Optimal cluster number ... 352
18.6.2 Elbow curve ... 352
18.6.3 Scaling data for k-means clustering ... 353
18.6.4 K-means clustering calculator ... 353
18.6.5 Key Features ... 354
19. Market Basket Analysis [Association Analysis] ... 355
19.1 What does association analysis do? ... 355

19.2 Market Basket Analysis Example ... 356
19.3 Interpreting the results of a Market basket analysis ... 358
19.3.1 Frequency ... 358
19.3.2 Support ... 359
19.3.3 Confidence ... 359
19.3.4 Lift ... 359
19.3.5 Market basket analysis and data mining ... 359
19.3.6 Critical note on the market basket analysis ... 360
20. Cronbach's Alpha ... 361
20.1 Latent variables ... 361
20.2 Assumptions for Cronbach's Alpha ... 363
20.3 Calculate Cronbach's Alpha ... 364
20.4 Example Cronbach's Alpha ... 364
20.5 Interpret Cronbach's Alpha ... 367
21. Cohen's Kappa ... 368
21.1 Cohen's Kappa Example ... 368
21.2 Inter-rater reliability ... 369
21.3 Use cases for Cohen's Kappa ... 369
21.4 Cohen's Kappa reliability and validity ... 370
21.5 Calculate Cohen's Kappa ... 371
21.6 Cohen's Kappa interpretation ... 374
21.7 Cohen's Kappa Standard Error (SE) ... 375
21.8 Calculating Standard Error of Cohen's Kappa ... 375
21.9 Interpreting Standard Error ... 375
21.10 Calculate Cohen's Kappa with DATAtab ... 376
22. Weighted Cohen's Kappa ... 377
22.1 Reliability and validity ... 378
22.2 Calculating weighted Cohen's Kappa ... 378
22.3 Calculate expected frequency ... 380
22.4 Calculate weighting matrix ... 381
22.5 Linear and quadratic weighting ... 382
22.6 Calculate weighted Kappa ... 383
22.7 Calculating weighted Cohen's Kappa with DATAtab ... 384
23. Fleiss Kappa ... 387
23.1 Fleiss Kappa Example ... 387
23.2 Fleiss Kappa with repeated measurement ... 388

23.3 Fleiss Kappa reliability and validity ... 389
23.4 Calculate Fleiss Kappa ... 390
23.5 Fleiss Kappa interpretation ... 393
23.6 Calculate Fleiss Kappa with DATAtab ... 394
24. Survival time analysis ... 396
24.1 Basics of survival time analysis ... 396
24.2 Use cases for survival time analysis ... 397
24.3 Example of survival time analysis ... 399
24.4 Censored data ... 400
24.5 Methods of survival time analysis ... 401
24.6 Kaplan-Meier Curve ... 402
24.6.1 Survival rate ... 402
24.6.2 Interpreting the Kaplan-Meier curve ... 404
24.6.3 Calculating the Kaplan-Meier curve ... 404
24.6.4 Drawing Kaplan Meier curve ... 406
24.6.5 Censored data ... 407
24.6.6 Comparing different groups ... 409
24.6.7 Kaplan-Meier curve assumptions ... 409
24.6.8 Create Kaplan Meier curve with DATAtab ... 410
24.7 Log Rank Test ... 411
24.7.1 Hypotheses in the Log Rank Test ... 415
24.7.2 Assumptions for the Log Rank Test ... 416
24.7.3 Calculate Log Rank Test ... 416
24.7.4 Calculate Log Rank Test with DATAtab ... 421
25. Cox Regression ... 424
25.1 Survival time analysis ... 424
25.1.1 Censoring ... 425
25.1.2 Cox Regression Example ... 426
25.1.3 Calculate Cox Regression with DATAtab ... 427
25.1.4 Interpretation of the Cox Regression ... 428
25.1.5 Assumptions of a Cox Regression ... 429
25.1.6 Calculate survival time analysis with DATAtab ... 431
26. z-Score ... 435
26.1 What is z-standardization? ... 435
26.2 Example of z-standardization ... 435
26.3 Calculating the z-score ... 440

26.4 Compare different data sets with the z-score 450
26.5 Assumptions z-standardization 451
References 454

List of Figures

Figure 1: Population and sample 20
Figure 2: Subareas of descriptive statistics 21
Figure 3: Task of inferential statistics 25
Figure 4: Methods of inferential statistics 26
Figure 5: Example inferential statistics 27
Figure 6: Scale or measurement levels 29
Figure 7: Sampling 33
Figure 8: Position dimensions 38
Figure 9: Representation of the median 41
Figure 10: Median for even and odd number of feature carriers 41
Figure 11: Mean and median in comparison 42
Figure 12: Measures of dispersion at a glance 47
Figure 13: Calculation of the standard deviation 49
Figure 14: Representation of the span 51
Figure 15: Illustration of the interquartile range 53
Figure 16: Example of a frequency table 56
Figure 17: Percentages and valid percentages 57
Figure 18: Frequency of car brands 61
Figure 19: Example of a crosstab 62
Figure 20: Creation of a crosstab 63
Figure 21: The most popular diagrams at a glance 66
Figure 22: Horizontal and vertical bar charts 67
Figure 23: Bar chart for frequencies 68
Figure 24: Bar chart falls in the hospital 69
Figure 25: Grouped bar charts 69
Figure 26: Bar chart for mean values 70
Figure 27: Error bar 71
Figure 28: Example of a bar chart 72
Figure 29: Nested bar chart 72
Figure 30: Example of a histogram 73
Figure 31: Example of a histogram 75
Figure 32: Example of a scatter plot 76
Figure 33: Interrelationships in the scatter diagram 77
Figure 34: Boxplot example 78
Figure 35: Interpret boxplot 80
Figure 36: Interpret boxplot part 2 81
Figure 37: Interpret boxplot part 3 82
Figure 38: Interpret boxplot part 4 83
Figure 39: Bland-Altman Plot 84
Figure 40: Create charts with DATAtab 88
Figure 41: Aim of inferential statistics 90
Figure 42: Properties of hypotheses 91
Figure 43: Hypotheses in the research process 92
Figure 44: Difference hypotheses 94
Figure 45: Correlation hypotheses 95

Figure 46: Directional and undirectional hypotheses 96
Figure 47: One-sided and two-sided testing 96
Figure 48: Logic of statistical inference 99
Figure 49: Hypothesis testing and research process 100
Figure 50: Uncertainty in hypothesis testing 101
Figure 51: Types of errors in hypothesis tests 104
Figure 52: Selection of the scale level with DATAtab 106
Figure 53: Overview Hypothesis tests 107
Figure 54: Example p-value 109
Figure 55: The interpretation of the p-value 110
Figure 56: The t-distribution and chi-square distribution 114
Figure 57: The Levene test of variance homogeneity 118
Figure 58: Graphical test of variance homogeneity 120
Figure 59: Explanation of the Levene test 121
Figure 60: Tests for normal distribution 123
Figure 61: Disadvantage of analytical tests for normal distribution 124
Figure 62: Histogram with normal distribution curve 125
Figure 63: Q-Q plot for testing normal distribution with DATAtab 126
Figure 64: Results of the test for normal distribution with DATAtab 126
Figure 65: Multicollinearity 128
Figure 66: The 3 variants of the t-test 132
Figure 67: One-sided and two-sided t-test 133
Figure 68: t-test statistic 135
Figure 69: t-test types 141
Figure 70: Mean difference 142
Figure 71: Calculation of the t-test for independent samples. 147
Figure 72: Calculation of the t-value 148
Figure 73: Boxplot showing the t-test results. 154
Figure 74: Forms of the t-test 157
Figure 75: t-test for dependent and paired samples, respectively. 158
Figure 76: t-test and Mann-Whitney U-test 167
Figure 77: Assumptions of the U-test 168
Figure 78: Calculate rank sums 169
Figure 79: Calculation of the Wilcoxon test 181
Figure 80: Example of categorical variables 194
Figure 81: Use of the chi-square test 194
Figure 82: From questionnaire to crosstab 195
Figure 83: Types of analysis of variance 208
Figure 84 Variance elucidation of the ANOVA 211
Figure 85: Measurement repetitions 217
Figure 86: Independent and dependent sample 218
Figure 87: Example independet and dependet sample 219
Figure 88: Example null and alternative hypothesis 220
Figure 89: Two-way ANOVA 227
Figure 90: One-factorial vs. two-factorial ANOVA 230
Figure 91: Interaction effect part 1 240
Figure 92: Interaction effect part 2 241
Figure 93: Two-way ANOVA with measurement repetition 242

Figure 94: Example two-way ANOVA with measurement repetition 244
Figure 95: Analysis of variance and Kruskal-Wallis test 250
Figure 96: Assumptions of the Kurskal-Wallis test 252
Figure 97: Scatter plot and correlation 260
Figure 98: Strength and direction of the correlation coefficients 264
Figure 99: Partial correlation 268
Figure 100: Bogus correlation storks and birth rate 269
Figure 101: Question of the regression 275
Figure 102: Types of regression 279
Figure 103: Simple and multiple linear regression 281
Figure 104: Scatter plot to show the correlation. 282
Figure 105: Representation of the regression line 283
Figure 106: Dichotomous variables in logistic regression 298
Figure 107: Factors influencing a disease in the regression model 298
Figure 108: Limits of linear regression 300
Figure 109: The logistic function 301
Figure 110: Approximation of the logistic function 302
Figure 111: Likelihood function 303
Figure 112: Example binary logistic regression 311
Figure 113: Confidence Interval 320
Figure 114: Calculating confidence intervals 323
Figure 115: Calculating confidence intervals 324
Figure 116: Factor Analysis 326
Figure 117: Basics of factor analysis 328
Figure 118: Scree-Test 331
Figure 119: Cluster Analysis example 335
Figure 120: Cluster Analysis steps 337
Figure 121: Euclidean Distance 338
Figure 122: Market basket analysis 1 355
Figure 123: Market basket analysis 2 356
Figure 124: Market basket analysis 3 356
Figure 125: Scale with 4 items 361
Figure 126: Latent variable 362
Figure 127: One and multiple latent variable(s) 363
Figure 128: Cohen’s Kappa example 369
Figure 127: Calculation of Cohen’s Kappa example 371
Figure 130: Cohen’s Kappa calculation with DATAtab 376
Figure 128: Calculating Fleiss Kappa with DATAtab 395
Figure 132: Survival time analysis 396
Figure 133: Survival time analysis use case 397
Figure 131: Survival time analysis example 399
Figure 132: Kaplan Meier Curve 406
Figure 133: Log Rank Test 411
Figure 134: z-standardization example 435
Figure 135: Calculating th z-standardization 440

List of tables

Table 1: Table of t-values 136
Table 2: Strength of the correlation 259
Table 5: Results of the linear regression 295

Introduction

This book provides an overview of key topics in statistics. The main methods and features of descriptive and inferential statistics are described and illustrated with graphics. In addition, the book offers step-by-step explanations for data analysis with the statistics software DATAtab.

The aim is to present the background of the statistical methods and their implementation in DATAtab in an easy-to-understand and clear way.

We start with the basics of descriptive and inferential statistics, their differences, and applications. This is followed by an introduction to the central basic concepts of statistics. The focus is on the concepts of variable or characteristic, scale or level of measurement, sample, population, and complete survey.

We then move on to statistical procedures for testing for differences in more than two groups and look at various forms of analysis of variance. Building on this, we look at statistical techniques for testing correlations and explore the field of correlation analysis and partial correlations.

Finally, we look at regression, with examples of linear and logistic regression. We conclude with structure discovery methods such as factor analysis and k-means cluster analysis.

Throughout the book, we try to explain why each term or method is important, at what point in the research process it is relevant, and what questions it can be used to answer.

We hope you enjoy reading and learning!

1. Descriptive statistics and inferential statistics

Descriptive statistics and inferential statistics, along with exploratory statistics, are the main areas of statistics. Descriptive statistics provides tools to describe a sample. Starting from the sample, inferential statistics can now be used to make a statement about the population.

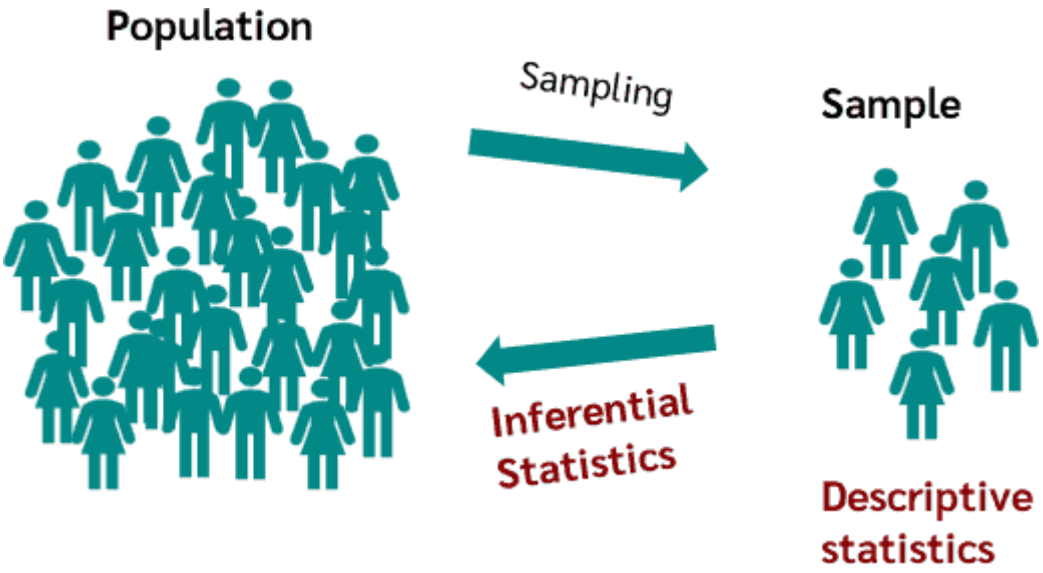


Figure 1: Population and sample

1.1 Subareas of statistics

One main area of statistics is to make a statement about a population. In most cases it is not possible to get all data of the population, so a sample is taken. This sample can now be described using descriptive statistics, e.g. what the mean value is and how strongly the sample scatters.

But this is not yet a statement about the population, that is the task of the inferential statistics. The inferential statistics takes a sample from the population, in order to make inferences about the population with this sample. So, the goal of inferential statistics is to infer the unknown parameters of the population from the known parameters of a sample.

Therefore, inferential statistics try to infer conclusions that go beyond the immediate data, unlike descriptive statistics. To achieve this, hypothesis tests such as the t-test or analysis of variance are used in inferential statistics.

1.2 Descriptive statistics

After collecting data, one of the first things to do is to graph the data, calculate the mean and get an overview of the **distributions of the data**. This is the task of descriptive statistics.

Thus, the goal of descriptive statistics is to gain an overview of the distribution of data sets. Descriptive statistics helps to **describe and illustrate** data sets.

Definition: The term descriptive statistics covers statistical methods for describing data using **statistical characteristics, charts, graphics or tables**.

It is important here that only the properties of the respective sample are described and evaluated. However, no conclusions are drawn about other points in time or the population. This is the task of inferential statistics or concluding statistics.

The various sub-areas of descriptive statistics can be summarized as follows:

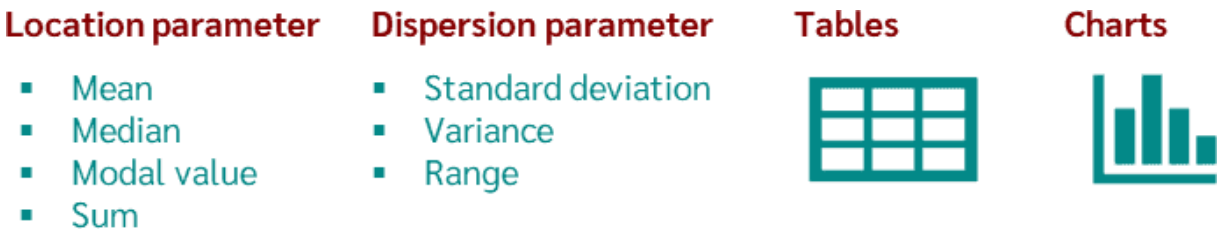


Figure 2: Subareas of descriptive statistics

Depending on which question and which measurement scale is available, different key figures, tables and graphics are used for evaluation. The best known of these are:

- **Location parameter:** Mean value, median, mode, sum
- **Dispersion parameter:** Standard deviation, variance, range
- **Tables:** Absolute, relative and cumulative Frequencies
- **Charts:** Histograms, bar charts, box plots, scatter charts, matrix plots

The first group of descriptive statistics are **location parameter** like the mean and mode. They are used to express a central tendency of the data set. They therefore describe where the center of a sample is located or where most of the sample is.

The second group are **measures of dispersion**. They provide information about how much the values of a variable in a sample differ from each other. Measures of dispersion can therefore describe how strongly the values of a variable deviate from the mean value: Are the values rather close together, i.e. are they similar, or are they far apart and thus differ greatly? A classic example is the standard deviation.

Which measures of location or dispersion are suitable for describing the data depends on the respective scales of measurement of the variable. Here, a distinction can be made between metric, ordinal and nominal scales of measurement.

Finally, a large area of descriptive statistics is diagrams such as the bar chart, the pie chart, or the histogram.

That’s how it works with DATAtab:

With DATAtab you can create diagrams directly in your web browser. The following example shows the steps involved.

Example: A sample of 10 male basketball players is drawn and their height is measured in meters. To get started, go to datatab.net and copy the data below into the Statistics Calculator table. Then you click on “Descriptive Statistics” in the calculator and select the variable "Height".

DATAtab will now give you the following table with descriptive statistics on the height of the players. The table shows the relevant dispersion and location parameter.

Statistics

	Height
Mean value	1.67
Median	1.655
Fashion	1.64
Total	16.7
Standard deviation	0.066
Variance	0.004
Minimum	1.55
Maximum	1.78
Range	0.23

Players	Height
1	1.62
2	1.72
3	1.55
4	1.7
5	1.78
6	1.65
7	1.64
8	1.64
9	1.66
10	1.74

1.3 Inferential Statistics

What's inferential statistics? In contrast to descriptive statistics, inferential statistics want to make a statement about the population. However, since it is almost impossible in most cases to survey the entire population, a sample is used, i.e. a small data set originating from the population.

With this sample a statement about the population can be made. An example would be if a sample of 1,000 citizens is taken from the population of all Canadian citizens.

Figure 3: Task of inferential statistics

Depending on which statement is to be made about the population or which question is to be answered about the population, different statistical methods or hypothesis tests are used.

The best known are the hypothesis tests with which a group difference can be tested, such as the t-test, the chi-square test or the analysis of variance.

Then there are the hypothesis tests with which a correlation of variables can be tested, such as correlation analysis and regression.

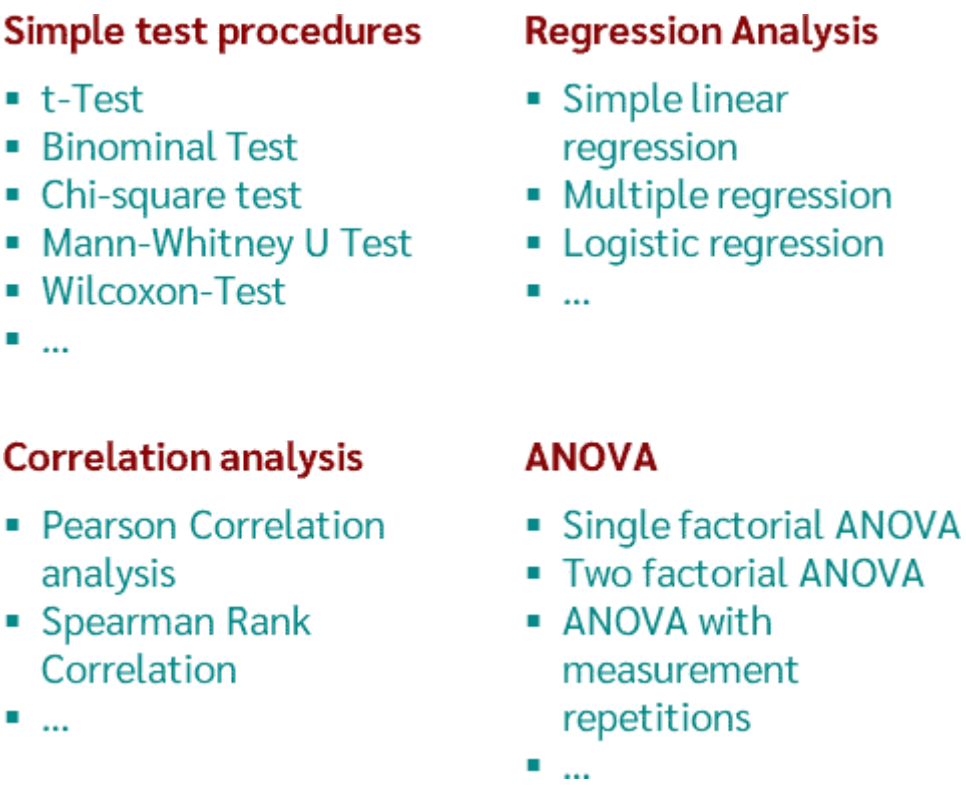


Figure 4: Methods of inferential statistics

Inferential statistics definition

Inferential statistics is a branch of statistics that uses various analytical tools to draw conclusions about the population from sample data. For a given hypothesis about the population, inferential statistics uses a sample and gives an indication of the validity of the hypothesis based on the sample collected.

Example inferential statistics

In the example above, a sample of 10 basketball players was drawn and then exactly this sample was described, this is the task of descriptive statistics. If you want to make a statement about the population you need the inferential statistics. For example, it could be of interest if basketball players are larger than the average male population. To test this hypothesis a t-Test is calculated, the t-test compares the sample mean with the mean of the population.

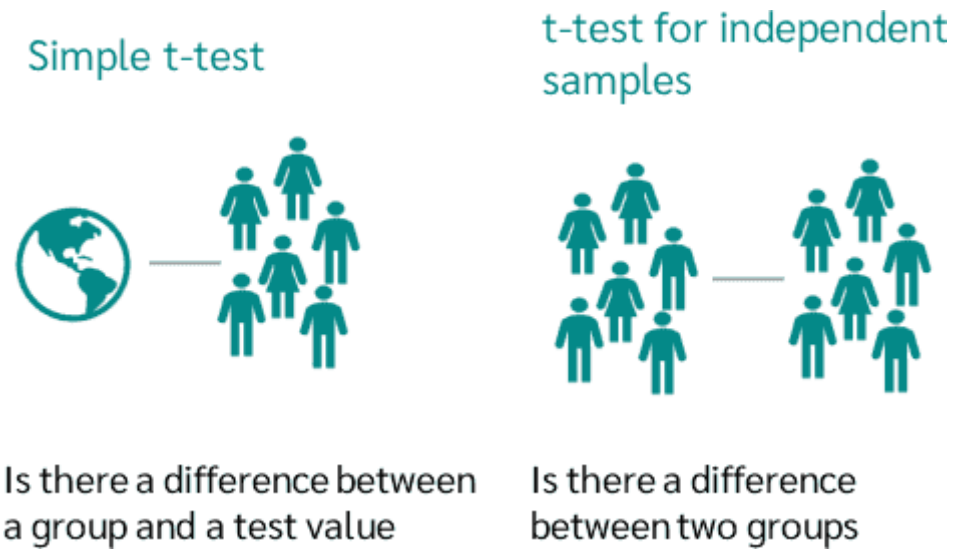

Simple t-test
t-test for independent samples
Is there a difference between a group and a test value
Is there a difference between two groups

Figure 5: Example inferential statistics

Furthermore, the question could arise whether basketball players are larger than football players. For this purpose, a sample of football players is drawn, and then the mean value of the basketball players can be compared with the mean value of the football players using an independent t-test. Now a statement can be made, for example, whether basketball players are larger than football players in the population or not.

Since this statement is only made based on the samples and it can also be pure coincidence that the basketball players are larger in exactly this sample, the statement can only be confirmed or re-submitted with a certain probability. In the table below you will find an overview of the most common inferential statistical methods used to make a statement about the population:

	Scales of Measurement		
	Nominal scale	**Ordinal scale**	**Interval scale**
One sample			
	Binominal Test		t-test for one sample
Two samples			
independent	Chi-square test	Median test	t-test for homogeneous variances
		Mann-Whitney U Test	t test for heterogeneous variances
dependent	McNemar-Test	Wilcoxon-Test	Paired t test
> two samples			
independent	kxm-Felder-Chi-Square Test	H-Test Kruskal & Wallis	ANOVA
dependent	Cochran-Test	Friedman-Test	ANOVA for repeated measurements

That’s how it works with DATAtab:

- DATAtab's Hypothesis Test Calculator allows you to calculate the various tests in the field of inferential statistics easily and directly online in your browser.
- You will find instructions on how to do this directly after the explanation of each statistical test in this book.

2. Level of measurment

One of the most important properties of variables is the level of measurement, also called scales of measurement. The measurement scale is important because it determines the permissible arithmetic operations and thus specifies the possible statistical tests. The higher the level of measurement, the more comparative statements and arithmetic operations are possible.

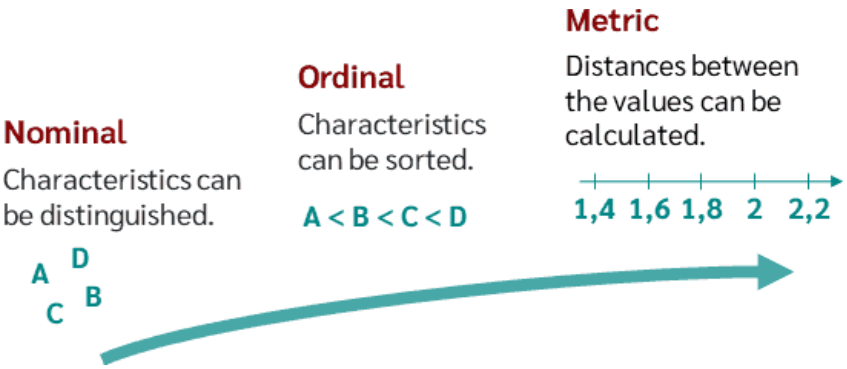


Figure 6: Scale or measurement levels

The **level of measurement** of a variable can be either **nominal, ordinal or metric.** In a nutshell: For nominal variables the values can be differentiated, for ordinal variables the values can be sorted and for metric scale level the distances between the values can be calculated. Nominal and ordinal variables are also called categorical variables.

2.1 Nominal variables

The nominal scale is the **lowest scale level** in statistics and thus has the lowest information content. Possible expressions of the variables can be distinguished, but a meaningful order is not possible. If there are only two expressions, such as in the case of gender (male and female), we also speak of **dichotomous** or **binary** variables.

- Only relations "equal", "unequal" possible
- No logical ranking of the categories
- The order of the answer categories is interchangeable

- Nominal characteristics with only two expressions are also called "binary" or "dichotomous".

Examples:

Gender	Marital status	You read the newspaper:
1 = male	1 = single	1 = The Washington Post
2 = female	2 = married	2 = The New York Times
	3 = divorced	2 = USA Today
	4 = widowed	...

2.2 Ordinal variables

The ordinal level of measurement is the next higher level, it contains nominal information, only with the difference that a ranking can be formed, therefore the term ranking scale is often used. In these cases, however, the distances between the values are not interpretable, so it is not possible to make a statement about the absolute distance between two values. A classic representative of the ordinal scale are school grades, here a ranking can be formed, but it cannot be said that the distance between A and B is the same as the distance between B and C.

- Next higher scale of measurement
- "equal" and "unequal" or "greater" and "smaller" can be determined
- There is a logical hierarchy of categories
- The distances between the numerical values are not equal, i.e. cannot be interpreted

Examples:

Frequency of television:	The government is doing a good job:
1 = daily	1 = agree with
2 = several times a week	2 = undecided
3 = less frequently	3 = disagree with
4 = never	

2.3 Categorical variables

Variables that have a nominal scale or an ordinal scale are also called categorical variables. In other words, categorical is an umbrella term for variables scaled nominally and ordinally.

Categorical variables can have a limited and usually fixed number of expressions, e.g. country with Germany, Austria, ... or gender with female and male. It is important, however, that it must be a finite number of categories or groups. The different categories can have a ranking, but do not have to.

2.4 Metric variables

Metric variables have the **highest possible level of measurement**. With a metric level of measurement, the characteristic values can be compared and sorted and distances between the values can be calculated. Examples would be the weight and age of subjects.

- Highest level of measurement
- Creation of rankings possible
- "equal" and "unequal", "greater" and "smaller" can also be determined
- Differences and sums can be formed meaningfully

Examples:

Income	Weight	Age	Electricity consumption
1820 $	81 kg	18 years	520 kWh
3200 $	70 kg	27 years	470 kWh
800 $	68 kg	64 years	340 kWh
...	...	...	...

2.5 Ratio scale and interval scale

The metric level of measurement can be further subdivided into interval scale and ratio scale. As the name suggests, the values of the ratio scale can be put into a ratio. Thus, a statement like the following can be made: "One value is twice as large as another". For this, an absolute zero must be available as a reference.

Example ratio scale:

The time of marathon runners is measured. Here the statement can be made that the fastest runner is twice as fast as the last runner. This is possible because there is an absolute zero point at the beginning of the marathon where all runners start from zero.

Example interval scale:

If, however, the stopwatch is forgotten to start at the start of the marathon and only the differences are measured starting from the fastest runner, the runners cannot be put in proportion. In this case it can be said how big the interval between the runners is (e.g. runner A is 22 minutes faster than runner B), but it cannot be said that runner A ran 20 percent faster than runner B.

The classic example is the temperature indication in degrees Celsius and Kelvin. The zero point of the Kelvin temperature scale is absolute zero, therefore it is a ratio scale. At degrees Celsius the absolute zero point is -273.15 °C, therefore the value zero on the degree Celsius scale cannot be assumed as natural zero and therefore it is an interval scale.

3. Sampling

In the following chapter we will discuss central concepts of sampling theory. You will learn what a population is and how to select elements from it. The way you draw your sample influences which statistical methods are useful for your topic and what statements you can make.

When planning an empirical study (e.g. a survey), sampling is very important. Before you start collecting your data, you need to decide how you will select the people who will take part in your study.

3.1 Full or total survey vs. random sample

The first question to ask is whether you need to draw a sample, or whether you will conduct a full or total survey. In a complete or total survey you collect data on all persons of the population or you already have data on all these persons. This is often the case, for example, when you are working with administrative data (e.g., grade lists of all students at a university) or have user data available (e.g., sales figures in an online store). In practice, full or total surveys are usually difficult to implement because they are expensive and time-consuming. Therefore, if you are writing your bachelor's or master's thesis and want to do a survey, you will most likely have to define a sample. (Häder 2010: p. 139)

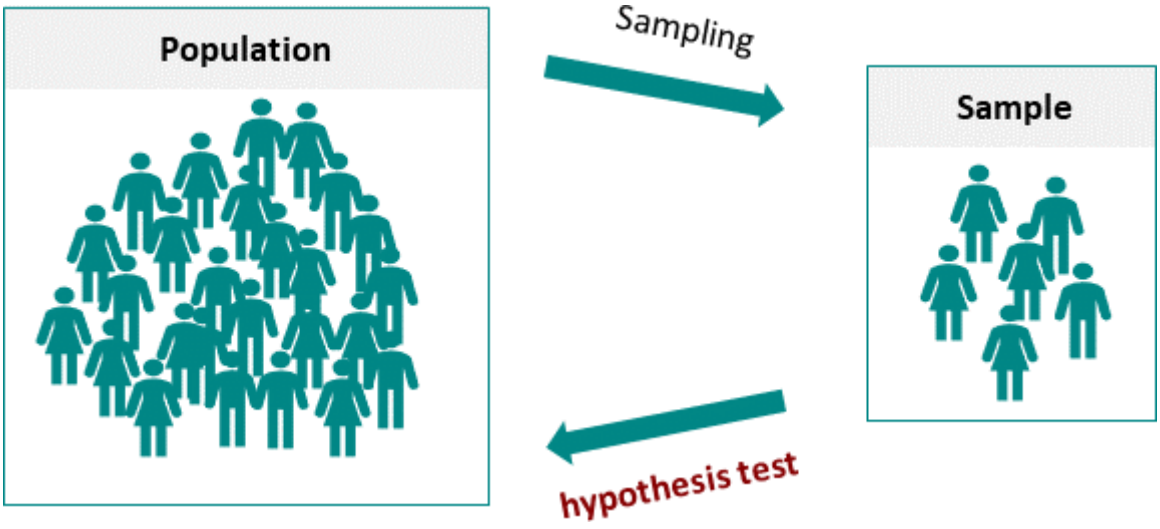


Figure 7: Sampling

3.2 Population and sample

As explained above, in a complete or total survey you are working with everyone in the population. This means that you have data on the whole population. What is the population? The population is all the elements that are of interest for the research. For example, it could be all the people you want to find out about through a survey. A sample is a selection from the entire group of elements, i.e. a selection from the population (Häder 2010, p. 141).

3.3 Types of sampling?

There are several ways in which you can draw a sample. Thus, a sampling procedure defines the way and the steps you use to select the elements from your population. Three main groups of sampling procedures can be distinguished (cf. Diekmann 2008, p. 378):

- Probability selection
- Deliberate selection
- Arbitrary selection

In the next section we will look at these three forms in more detail and discuss examples. We will then look at how to deal with sampling in an online survey and how to describe sampling in a bachelor or master thesis.

3.4 Probability selection

When you perform probability selection, you get a random sample. For example, you need a list of all elements of the population from which you randomly select individuals. It is important that each element of the population has the same probability of being included in the sample.

An example of this would be a random selection of households from the central population register of a city. Using a computer, you can then randomly select from this register, for example, a sample of 1000 addresses in the city. You then contact these households and ask them (or a selected member of the household) to participate in your survey.

Probability selection is often difficult to implement in practice, however, because in many cases there is no list of the population, or the procedure of this selection is too elaborate to be implemented in smaller empirical studies.

You can do probability selection in a single-stage or multistage way. Single-stage would be that you select your items in one step. Multi-step means that you make your selection in several steps. For example, in the first stage you select 50 municipalities of a state and in the second stage you select 50 addresses per municipality from these municipalities. (Diekmann 2008: pp. 380-386)

3.5 Deliberate selection

Deliberate selection is based on certain criteria, and it is based on the distribution of characteristics in the population. Quota sampling is a well-known example of deliberate selection. To draw a quota sample, you look at how certain characteristics (e.g., age and gender) are distributed in your population. Quota characteristics can be gender, age, educational attainment, place of residence, different hierarchies in a company, length of tenure with a company, etc.

For example, if you are doing a survey in retail and you see that in your population there are 40% younger women (up to 39 years), 30% older women (40 years and older), 20% younger men and 10% older men, you try to achieve this distribution in your sample as well.

Quota sampling is especially widespread in the field of market and opinion research and is also often implemented in the context of bachelor or master theses. This form of sampling is less time-consuming and cost-intensive and can also be implemented well in smaller empirical studies. However, an important prerequisite is that you have information about your population and know how certain characteristics (e.g. age, gender, etc.) are distributed there.

3.6 Arbitrary selection

The third group of sampling methods is arbitrary selection. Here you do not control the process of sample selection. Arbitrary selection is often used in experiments in psychology. Here you do not select your test persons purposefully, but it takes part, who can and would like to take part. (Diekmann 2008: p. 379).

3.7 Sample selection in online surveys

With online surveys, it is mostly more difficult to determine the sample selection in advance. In most cases, there is no list of the population from which a selection can be made. One possibility here is to repeatedly look at the already completed questionnaires during the course of the survey and check the distributions of quota characteristics. If, for example, you notice that older women are underrepresented, you can still actively approach or write to this target group. The goal is to get as close as possible to the quota plan.

3.8 Sample description in bachelor or master thesis

No matter which sampling method you choose, it is very important that you explain your approach well in your bachelor or master thesis and make it

comprehensible. It should be clear to the reader of your thesis what the population and sample of your study are and how you have selected them.

In your paper, you should answer the following questions:

- Who was selected and interviewed (sample, population)?
- Why were these people selected?
- How did you contact respondents?

4. Location parameter

In descriptive statistics, the **mean, median and mode** are location parameter (measures of central tendency). Based on data collected in a sample, the location parameter provide information about where the "center" of the distribution lies.

Measures of location can be used to summarize or describe a list of data with only one parameter. An example would be that the average duration of studies of sports students at the university XY is 11.1 semesters.

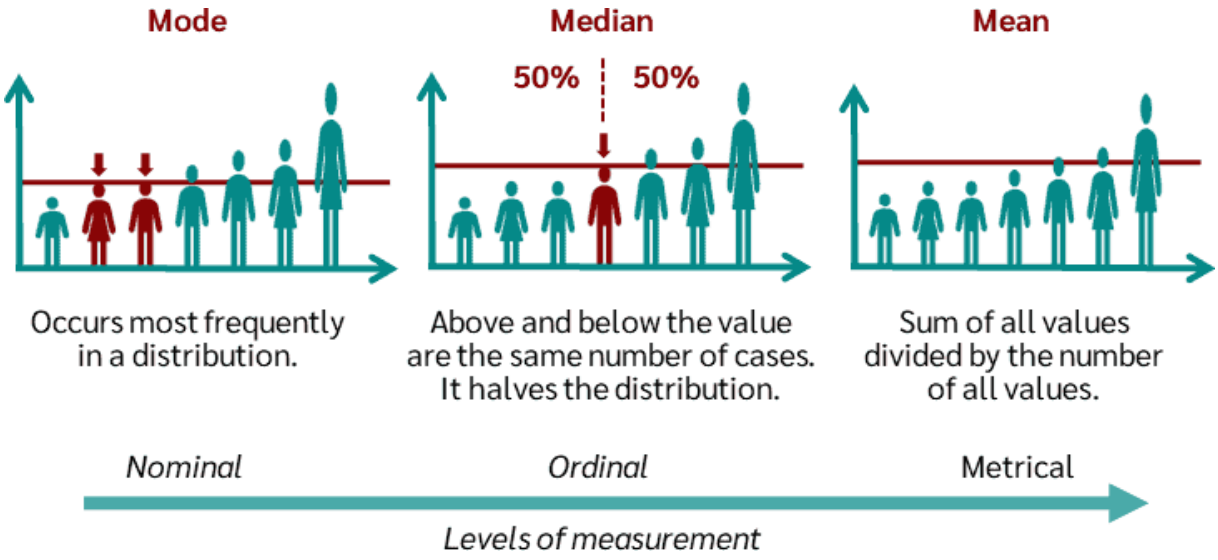


Figure 8: Position dimensions

Together with the dispersion parameters, the position parameters thus describe a distribution in statistics. The most commonly used position parameters are the mode, the median and the mean. Which location parameter is used depends on the scale level of the variable and the robustness against outliers.

4.1 Mean value (arithmetic mean)

The mean value can only be calculated for **metric variables**, i.e. when metric scale level is given. It indicates where the center of gravity of a distribution is to be found. In everyday life, it is also referred to as the **"average"**.

Definition: The **arithmetic mean** is the sum of all observations divided by their number n.

How can the mean value be calculated?

The mean value can be calculated by **adding all expressions of a variable** and finally **dividing the sum** by the **number of characteristic expressions.**

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Mean Value — Number of values — Value at the i-th position

Example: A group of 5 statistics students was asked how many cups of coffee they drink per week. The result is 21, 25, 10, 8 and 11 cups. The average is thus 15.

1	2	3	4	5
21	25	10	8	11

$$\frac{21 + 25 + 10 + 8 + 11}{5} = 15$$

Tip: With DATAtab's statistics calculator you can easily calculate the mean value or the desired position parameter for your data.

4.2 Geometric mean and quadratic mean

When talking about mean or average, mostly the arithmetic mean is meant, but there are also other types of mean values.

Other mean values are, for example, the geometric mean and the quadratic mean also called Root Mean Square (RMS).

Root Mean Square

$$\bar{x}_{RSM} = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}$$

Geometric Mean

$$\bar{x}_{geom} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Geometric mean: If there are n positive numbers, the geometric mean is the nth root of the product of the n values.

Root Square Mean: The root square mean is obtained by dividing the sum of the squares by the number of values and taking the square root.

4.3 Median

If the measured values of a variable are ordered by size, the value in the middle is the median. The median is therefore the "middle value" of a distribution. It leads to a division of the series into two parts: One half is smaller and one larger than the median.

Since the data are ordered for the calculation of the median, the variables must have an ordinal or metric scale level.

Definition: For an ordered series, the **median** is the value that divides the series into an equal lower and upper range.

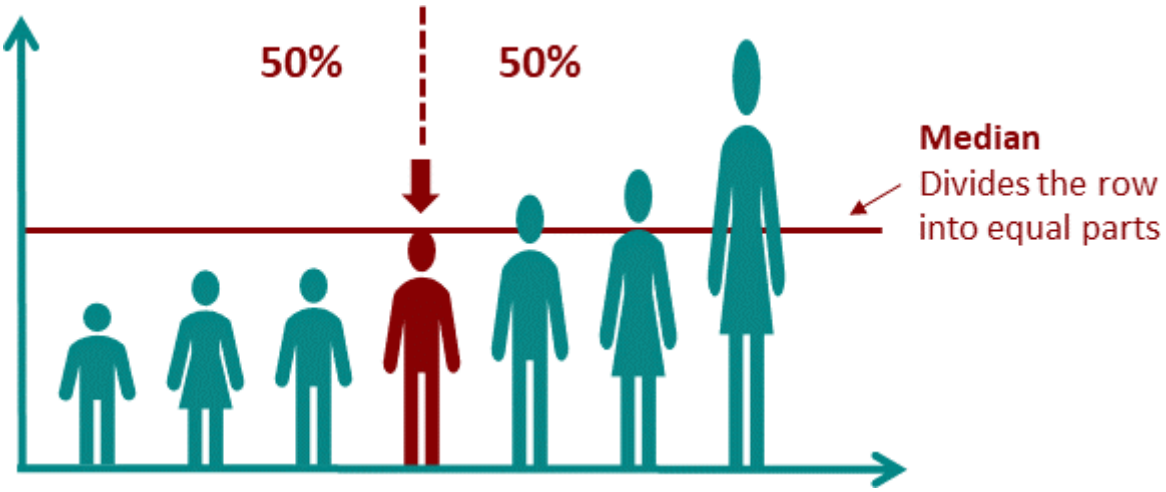


Figure 9: Representation of the median

For the **median** to be calculated, the variable must be **ordinally scaled. Ordinally scaled** means that there is a **ranking** between the values of a variable. This applies, for example, to school grades, height, or salary. In the case of a birthplace variable, however, it is not possible to establish a ranking, and therefore the median cannot be calculated here.

If there is an **odd number of** characteristic carriers (persons) and thus an odd number of expressions, then the median is a value that actually occurs.

If there is an **even number of** feature carriers (persons), the two middle expressions are added, and their sum is divided by two.

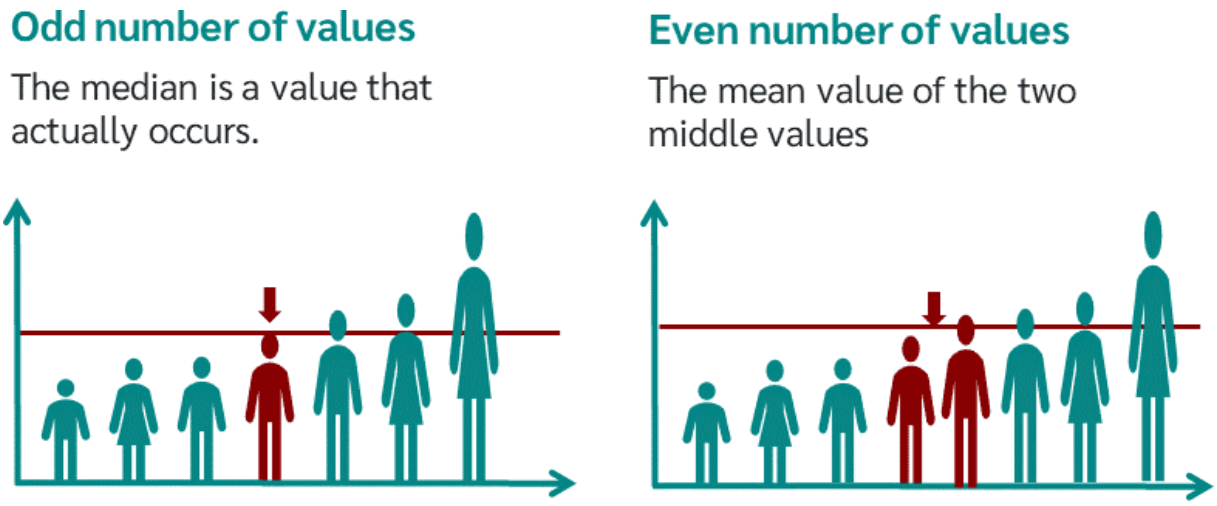

Odd number of values
The median is a value that actually occurs.
Even number of values
The mean value of the two middle values

Figure 10: Median for even and odd number of feature carriers

4.4 Mean and median in comparison

Compared to the mean, the median is much more robust against scattering. An outlier usually has no influence on the median, but it has a more or less large influence on the mean.

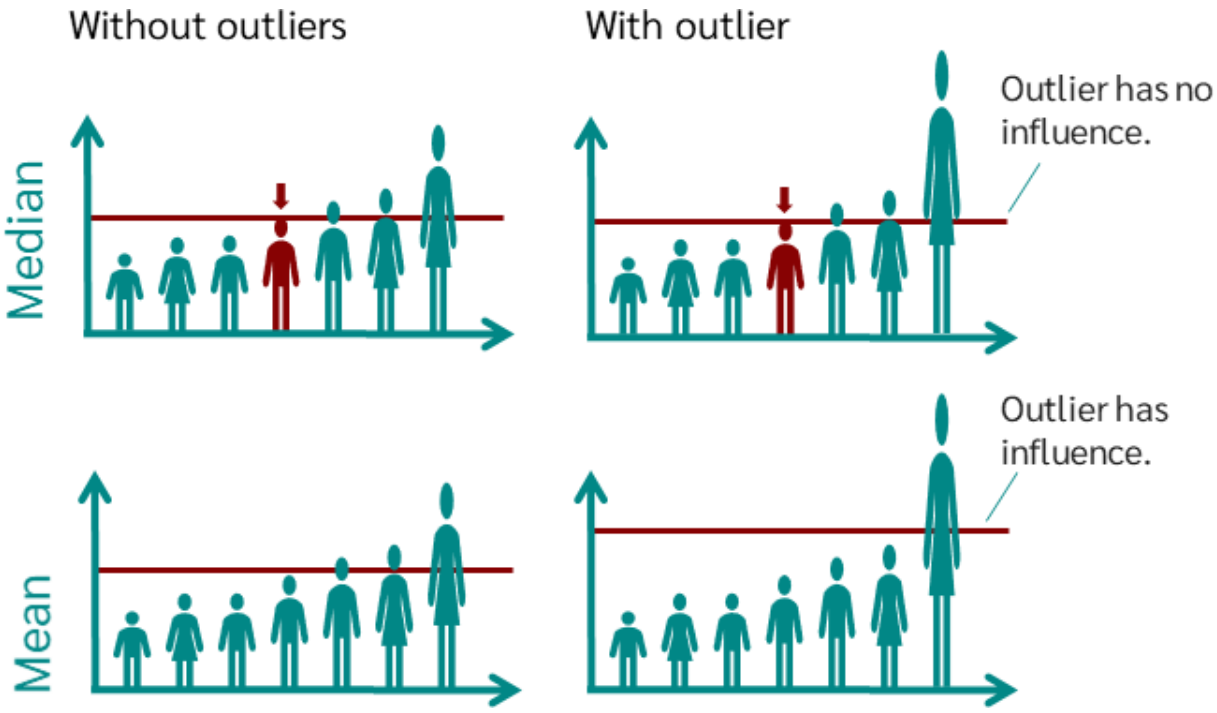


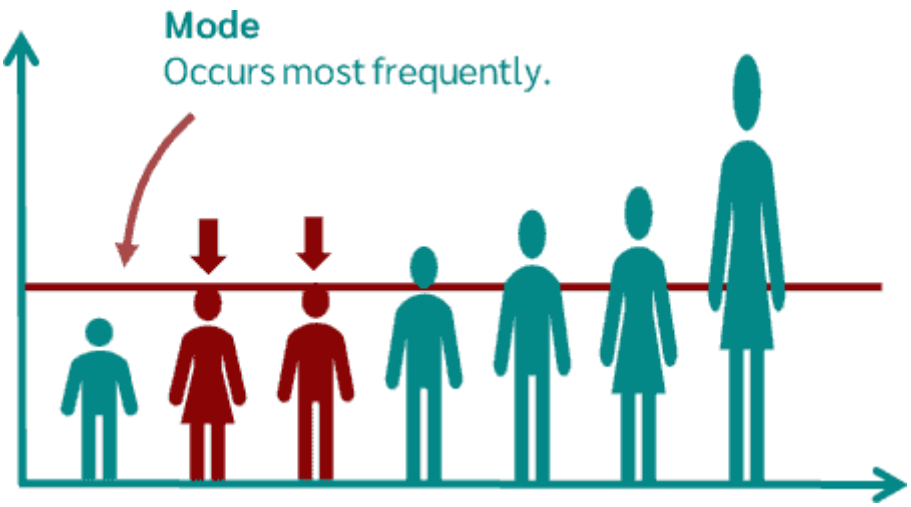
Figure 11: Mean and median in comparison

4.5 Mode (Modal value)

The mode is the most common value. The mode is therefore the most frequent value in a distribution and corresponds to the highest value in the distribution. It is therefore the value that is "typical" for a distribution.

The mode can be used for both metric and categorical (nominal or ordinal) variables.

Definition: The mode is the value of a distribution that occurs most often.



Calculate mode

Example: In a sample of 70 managers from Berlin, 20 drive a Daimler, 25 a BMW, 10 a VW and 15 an Audi. The car brand BMW is the most common. Thus the mode is "BMW".

Car brand	Daimler	BMW	VW	Audi
Frequency	20	25	10	15

Therefore the mode can easily be read in a frequency table, it is the most frequent observed value.

Attention: There can also be several mode values. If two or more points occur with the greatest frequency, then there are several mode values. In this case one speaks then of a bimodal or multimodal distribution.

4.6 Advantage and disadvantage of the Mean, Median and Mode

If the distribution is symmetric, the mean and median are equal, and if the distribution is symmetric and unimodal, all three measures are equal. As a rule, however, the three measures have different values. Now, of course, the question is which of the measures of central tendency to use. Unfortunately, there is no clear rule for this, only a few decision aids.

Mean: The mean value is by far the most used. The disadvantages of the mean are that it is sensitive to outliers, the value does not have to exist in the data and for the interpretation to be meaningful, the data should have metric scale level.

Median: The great advantage of the median is that it is very robust against outliers and that the data only have to be scaled ordinally.

Mode: The mode is the value that occurs most frequently, which has the advantage that the value actually occurs. Furthermore, the mode can also be calculated for data that cannot be ordered and thus have a nominal scale level. The disadvantage is that the mode does not take into account the other existing data.

That’s how it works with DATAtab:

- To calculate location measures with DATAtab, simply open the statistics calculator and insert your data into the table.
- After selecting your variables, you can then calculate, for example, the mean, median and mode of your data.
- As an example, the score can be used in a statistics exam. To do this, copy the data into the statistics calculator, click on "Descriptive statistics" and select the variable "Score".

Student	Score
1	4
2	5
3	5
4	8
5	9
6	12
7	14
8	16
9	17
10	20

The result then looks like this:

	Score
Mean value	11
Median	10,5
Fashion	5

Here you will find a **brief summary** of the three position measures discussed so that you have the definitions at a glance:

Calculate mean:

The mean value is obtained by dividing the sum of all values by their number of values.

Calculate median:

Due to the even number of values, the median is obtained by adding the two middle values. The sum is then divided by two.

Calculate modus:

To obtain the mode, the frequency of occurrence of each value is counted. The value that occurs most often is the modal value. In this case, the value 5 is the only one that occurs twice, so the modal value in this example is 5.

5. Dispersion parameter

Standard deviation, variance and range are among the measures of dispersion (Measurement of Variability) in descriptive statistics. They are calculated to describe the **scatter of values** of a sample around a location parameter. Put simply, **dispersion parameters** are a measure of how much a sample fluctuates around a mean value.

So, position measures give you information about the center of your data, scatter measures give you information about how much your data scatters around that center.

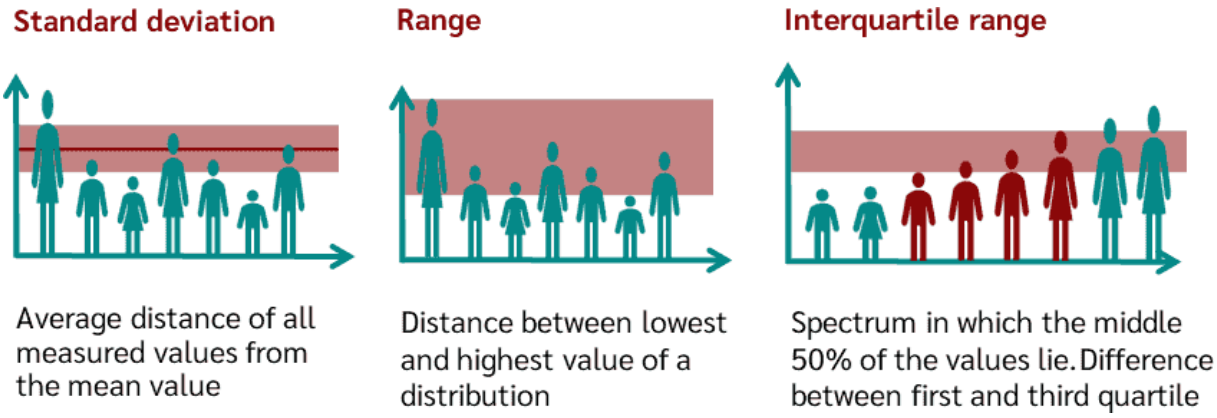

Standard deviation
Range
Interquartile range
Average distance of all measured values from the mean value
Distance between lowest and highest value of a distribution
Spectrum in which the middle 50% of the values lie. Difference between first and third quartile

Figure 12: Measures of dispersion at a glance

Measurement of Central Tendency give you the information about the centre of your data, dispersion measures give you the information how much your data is spread around this centre.

The most common measures of dispersion for metric variables are standard deviation and variance. These two measures relate each characteristic of a variable to the mean value and thus indicate how far the individual characteristics are scattered around the mean value.

5.1 Standard deviation

The standard deviation indicates the spread of a variable around its mean value. Thus, the standard deviation is the mean deviation (root mean square) of all measured values from the mean.

The standard deviation thus indicates how much the distribution of values scatters around the mean value. If the individual values scatter strongly around the mean value, a large standard deviation of the variable results. There are two slightly different equations for the calculation. On the one hand, the entire population can be used to calculate the standard deviation. On the other hand, it can also be calculated if only one sample is available. If all values of the population are available, the following results are obtained:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

n is the number of persons

x_i is the size of the individual

$\bar{x}$ is the mean value of all persons

Often, however, the data of the entire population are not available. Therefore, a sample is usually used to estimate the standard deviation of the population. In this case, the calculation results in

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

The difference between the two formulas is that one is divided by n and the other by n-1. It is customary to use **s** for **the standard deviation** of a sample and **σ** for the **standard deviation of the population**.

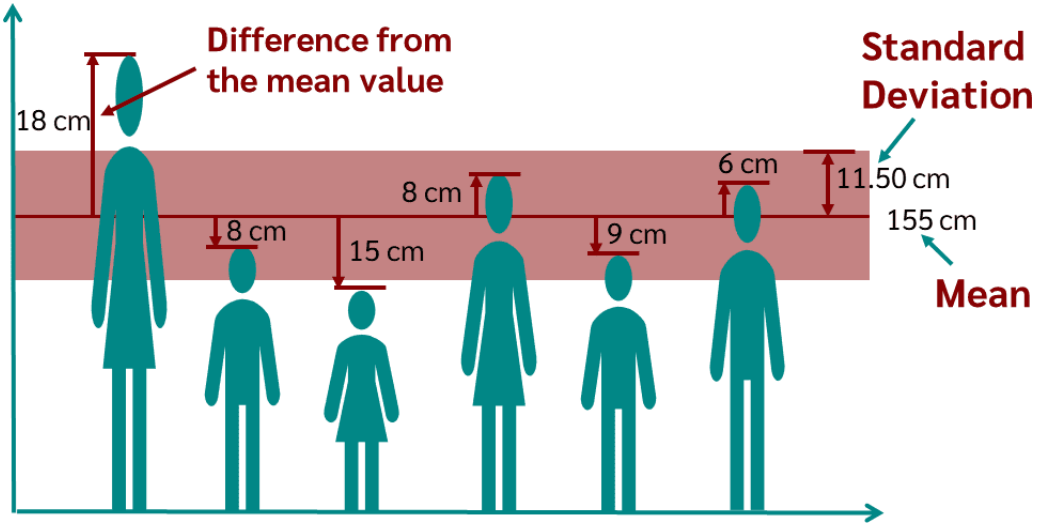


Figure 13: Calculation of the standard deviation

5.2 Variance

Just like the standard deviation, the variance measures the deviation from the mean. For the calculation of the variance, the sum of the squared variances is divided by the number of values.

$$Var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

The variance thus describes the squared average distance from the mean. Because the values are squared, the result has a different unit (the unit squared) than the original values. Therefore, it is difficult to relate the results.

5.3 Difference between variance and standard deviation

So, the difference between the dispersion parameter variance and standard deviation is that the standard deviation measures the average distance from the mean and the variance measures the squared average distance from the mean.

In other words, the variance is the squared standard deviation and the standard deviation is the root of the variance.

However, this squaring results in a key figure that is difficult to interpret, since the unit does not correspond to the original data.

For this reason, it is advisable to always use the standard deviation to describe a sample, as this makes interpretation easier.

Variance	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard deviation	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

5.4 Range

The range, also called span, is the distance between the minimum and maximum of a distribution, i.e. the distance between the smallest and the largest value.

For example, if the height of 7 people is queried and the largest value is 1.90m and the smallest is 1.50m, the span is calculated as 1.90m - 1.50m to 0.4m.

Definition Range: The range indicates the distance between the highest and the lowest value in a sample.

The range or span, often abbreviated with R, is therefore calculated by

$$R = x_{max} - x_{min}$$

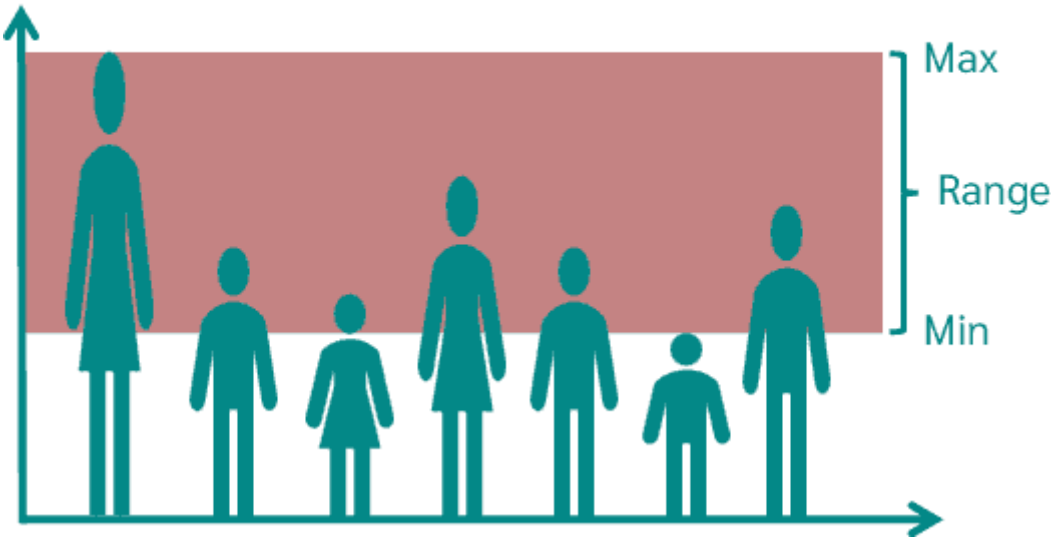
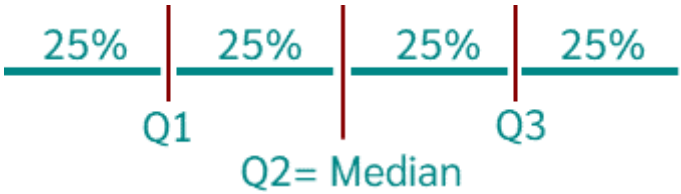


Figure 14: Representation of the span

5.5 Quartile

Quartiles divide your data into four parts, as equal as possible. For the calculation quartiles, the data must be sorted from the smallest to the largest value.

- Quartile (Q1): The middle value between the smallest value (minimum) and the median.
- Quartile (Q2): The median of the data, i.e. 50% of the values are smaller and 50% of the values are larger.
- Quartile (Q3): The middle value between the median and the largest value (maximum).



Thus, 25% of all values are below the lower quartile (Q1) and 75% are below the upper quartile (Q3).

5.6 Interquartile range

In contrast to the range in which 100% of all values lie, one often wants to know the range in which the middle 50% of all values lie. This scattering parameter is called interquartile range (IQR). The upper and lower 25% of the values are therefore not taken into account for the interquartile range.

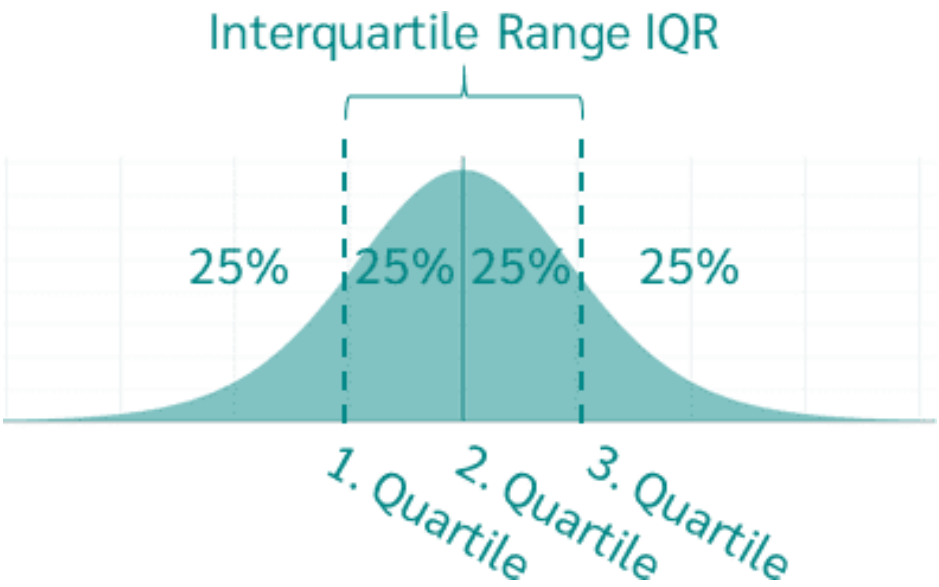


Figure 15: Illustration of the interquartile range

5.7 Example dispersion parameter

The calculation of range, variance and standard deviation shall now be illustrated by an example. For this purpose, the results of students in a statistics exam (scores) will be used.

Student	Score
1	4
2	5
3	5

Student	Score
4	8
5	9
6	12
7	14
8	16
9	17
10	20

Calculate Variance:

Varainz is the sum of the squared deviations from the mean value of all values divided by the number of values.

$$\frac{(11-4)^2+(11-5)^2+\ldots+(11-17)^2+(11-20)^2}{10-1}=31.778$$

Calculate standard deviation:

The standard deviation is the square root of the variance.

$$\sqrt{\frac{(11-4)^2+(11-5)^2+\ldots+(11-17)^2+(11-20)^2}{10-1}}=5.637$$

Calculate range:

The range is obtained by subtracting the smallest value from the largest value.

$$20-4=16$$

That’s how it works with DATAtab:

The calculator for descriptive statistics on DATAtab will give you the range, variance and standard deviation.

Copy the above data into the Online Statistics Calculator, click on Descriptive Statistics and select the Score variable. The result will look like this:

	Score
Standard deviation	5.637
Variance	31.778
Range	16

6. Frequency table

Frequency tables are created when you want to display the **absolute and relative frequencies** of the values of your variables or characteristics. Thus, a frequency table for your data shows you how often each characteristic occurs. A frequency table for the variable gender, for example, shows how often the characteristics male and female occur in the sample.

6.1 Absolute and relative frequencies

Absolute frequencies are those values that indicate how often the respective category of a variable occurs.

Relative frequencies, on the other hand, indicate how often the respective expressions occur in relation to all cases, and are therefore usually given as percentages.
Depending on the subject area and the question, the categories or characteristics can be, for example, persons, companies, locations, or households.

Frequency tables are often created to get an initial overview of data. Afterwards, the result can be displayed graphically in a bar chart.

	Frequency	%
Cake	2	22 %
Ice	4	44 %
Donut	3	33 %
Total	**9**	**100 %**

Figure 16: Example of a frequency table

6.2 Valid percent

It is particularly important to pay attention to **missing or invalid values** when **creating and interpreting frequency tables.** In the field of survey research, missing values are usually found where people have answered with "no answer," "Can't say," or "Don't know."

So that the statistics are not distorted by the missing values, you should indicate both **percentages and valid percentages** in the frequency table.

To calculate valid percent, you only need to divide the absolute frequencies of a characteristic by the valid cases.

If you asked 30 people in a survey what their favorite car brand was, and 7 ticked "Can't say", then there are 23 valid values.

If 5 people indicated Ford, then the valid percentages are 5/23 and 21.7%, respectively.

Figure 17: Percentages and valid percentages

6.3 Frequency table in statistics

Frequency tables mostly consist of the following columns:

- Absolute frequency
- Percentages (=relative frequency)
- Valid percent

The column of valid percentages is now the one that shows the relative frequencies of a characteristic in percent, but only considers those cases that have valid bets. Since missing values can always occur, it is advisable to also use this form of percentages.

Example for valid percent:

Let’s assume that a Sunday poll is taken, and the question is asked "Which party would you vote for if there were an election next Sunday?" it could turn out that there are still some undecided people.

In this case, both the percentages and the valid percentages would be interesting.

The percentage based on all values would therefore show how much support there is for a party in relation to all respondents, including the undecided.

The percentage of valid values, on the other hand, indicates the level of agreement among those who have already decided.

6.4 Example frequency table

The procedure of creating a frequency table will now be illustrated with an example:

Let’s assume that in a statistics course the participants were asked which brand of car they drive. The answers are displayed in the following table:

Student	**Car brand**
1	VW
2	
3	BMW
4	Skoda
5	Skoda
6	VW
7	BMW
8	Opel
9	Opel
10	Skoda
11	VW
12	Daimler

That's how it works with DATAtab:

- Simply copy the table into the online statistics software and select the variable car brand.
- Now you can choose which values you want to calculate.
- The result of the frequency table now looks like this:

Calculate:

☑ Frequency ☑ % ☑ Valid %

Copy Word Copy Excel

Car brand	Frequency	%	Valid %
VW	3	25%	27.27%
Skoda	3	25%	27.27%
BMW	2	16.67%	18.18%
Opel	2	16.67%	18.18%
Daimler	1	8.33%	9.09%
Total	11	91.67%	100%
Invalid	1	8.33%	
Total	12	100%	

The table shows the absolute frequencies, the relative percentages, and the valid percentages.

Finally, DATAtab automatically gives you a graphical visualization of the result, here in the form of a bar chart:

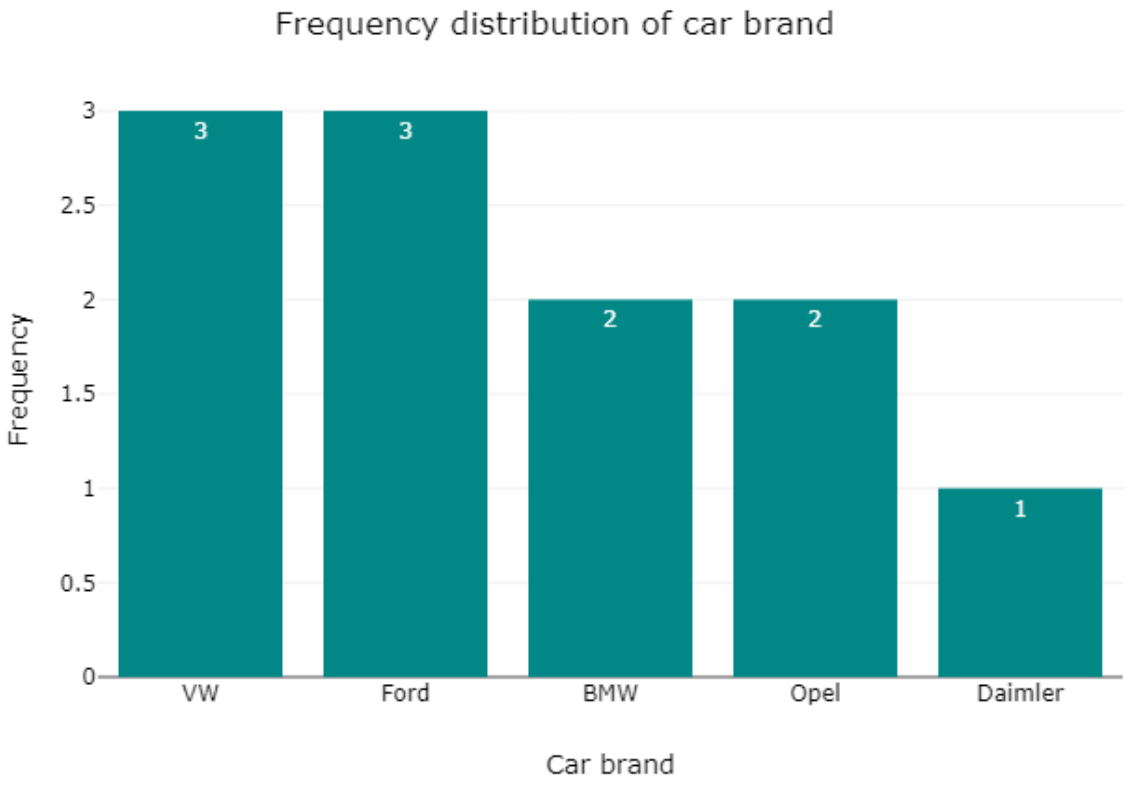


Figure 18: Frequency of car brands

If the variable is metrically scaled, a histogram is used to display the frequencies.

6.5 Frequency table in APA style

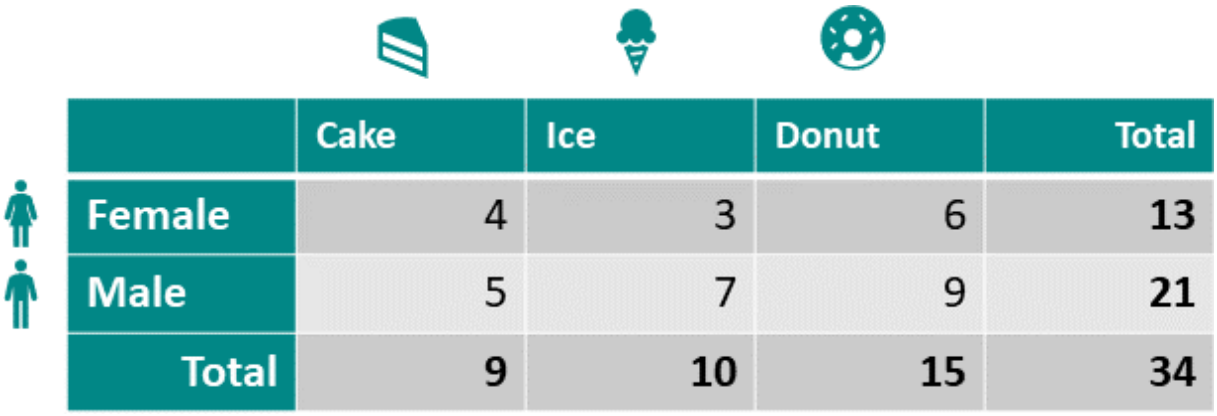
If you want to create a frequency table in **APA format**, you have to consider the following:

Font	Times New Roman, size 12
Labeling	All tables must be numbered in APA format
Margins	As few margins as possible should be used

7. Contingency table (Crosstab)

Crosstabs, also called contingency tables, are used in descriptive statistics to obtain an overview of two, usually categorical variables. In the crosstab, you can then read how often the combination of the values of two characteristics occurs.

The frequency is given either in **absolute or relative frequency**. Therefore, with a cross-tulationab in statistics, one gets an insight into how two variables are related.



	Cake	Ice	Donut	Total
Female	4	3	6	13
Male	5	7	9	21
Total	9	10	15	34

Figure 19: Example of a crosstab

7.1 Crosstabs in statistics

If two categorical variables are present, a crosstab is obtained by entering the values of the variables in a table. For the first variable, the values are plotted from left to right, for the second variable from top to bottom. The individual cells are then filled with either the **absolute** or the **relative frequency.**

Crosstabs and market research

Crosstabs are very often used in market research because they can be used to compare customers or products very well. For example, the following questions can be answered:

- Which insurance is preferred by which age group?
- Are the car brands different in the city and in the country?
- Which apple variety sells best in which season?

7.2 Interpretatation of crosstabs

How do you interpret a crosstab? A crosstab shows the frequencies of two variables.

In each cell of a crosstab, the frequencies of the characteristic combinations are entered; in the example above, "female and without a degree" occurs exactly 6 times.

Case	Gender	Highest level of education
1	Male	College
2	Female	Bachelor
3	Male	Without graduation
4	Male	Master
5	Female	Master
...	...	...

Female and without a degree occurs 6 times in the data

	Female	Male
Without graduation	6	7
College	13	16
Bachelor	16	15
Master	8	11
Total	43	49

Figure 20: Creation of a crosstab

Rows and columns in the crosstab

This means that the values of one variable are plotted in the rows and the values of the other variable are plotted in the columns. Usually, the independent variable is plotted in the columns and the dependent variable in the rows.

Absolute and relative frequencies for crosstabs

When creating a crosstab, either the absolute or the relative frequencies can be output:

- **Absolute frequencies** are those values that indicate how often the respective combination of two characteristic values occurs.
- **Relative frequencies,** on the other hand, indicate how often the respective combination of expressions occurs in relation to all cases; it is therefore usually expressed as a percentage.

7.3 Example crosstab

The creation of crosstabs will now be examined in more detail using an example.

In the example it is assumed that on a rainy day a student counts how many people "**with**" and how many "**without**" **umbrellas** come to the statistics lecture.

In addition, the student makes a note of the **gender** of the students.

Gender	With umbrella
female	yes
male	yes
female	yes
female	yes
male	yes
male	no
female	no
male	no
female	no
female	no

That's how it works with DATAtab:

- Just open the Statistics Calculator and copy the table above.
- Then you select the variables "gender" and "using an umbrella" in the "Descriptive statistics" section.
- DATAtab will then automatically create a crosstab for you.

The result is shown in the following **crosstab**. The cross-tabulation now contains the absolute frequencies of the respective feature combinations. This ist he result:

		With umbrella		
		yes	no	Total
Gender	female	5	7	12
	male	5	5	10
	Total	10	12	22

7.4 Testing a crosstab for significance

A crosstab can be used to examine whether there is a relationship between the two variables. However, since a crosstab is a descriptive statistic, a statement can only be made about the sample. If a statement is to be made about the population, the **chi-square test** is required.

8. Charts

In charts, data is presented graphically, so they are mainly used in statistics to get an initial overview of the collected data and to present information in an easily understandable way.

The most commonly used charts in statistics are bar charts, histograms, scatter plots, line plots, box plots or pie charts.

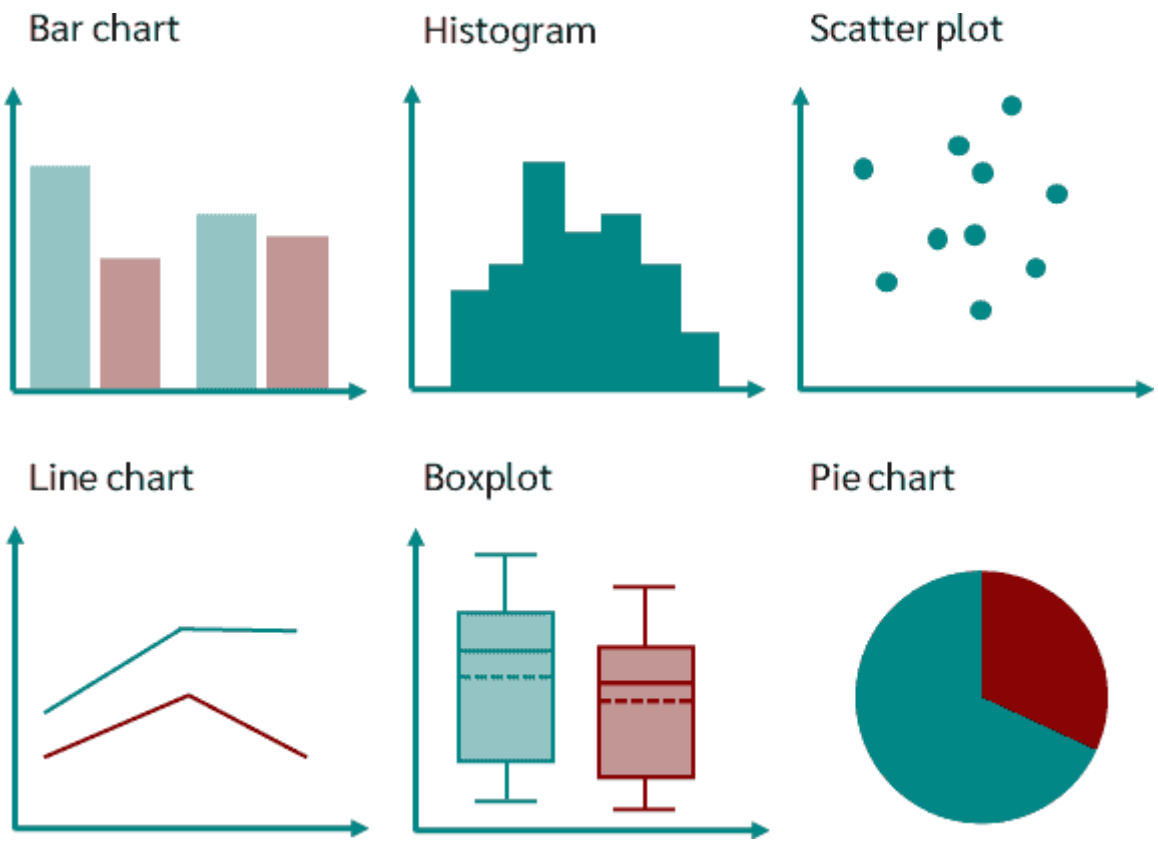


Figure 21: The most popular diagrams at a glance

8.1 Bar chart

Bar charts are probably the most commonly used charts in statistics. Bar charts are usually used to show the frequency of different categories, but also to visualize numerical data, such as sales figures or population statistics.

In a bar chart, the length of each bar is proportional to the value it represents. The bars are usually arranged horizontally or vertically.

- **Horizontal bar chart:** Frequencies are represented by horizontal bars and the y-axis plots the characteristic values.
- **Vertical bar chart:** Frequencies are represented by vertical bars and the characteristic values are plotted on the x-axis.

Figure 22: Horizontal and vertical bar charts

8.2 Bar chart for frequencies

In a bar chart, absolute and relative frequencies are displayed on a two-axis coordinate system.

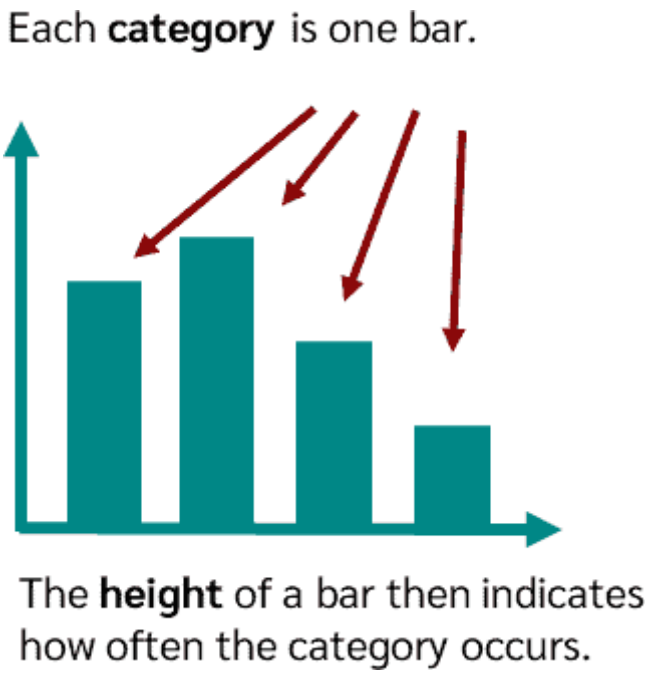


Figure 23: Bar chart for frequencies

- **Horizontal bar chart**: Frequencies are represented by horizontal bars and the y-axis plots the characteristic values.

- **Vertical bar chart**: Frequencies are represented by vertical bars and the x-axis plots the characteristic values.

Due to the simplicity of bar charts, they are often created in descriptive statistics. They provide a very quick overview of the ranking and frequencies of characteristic values.

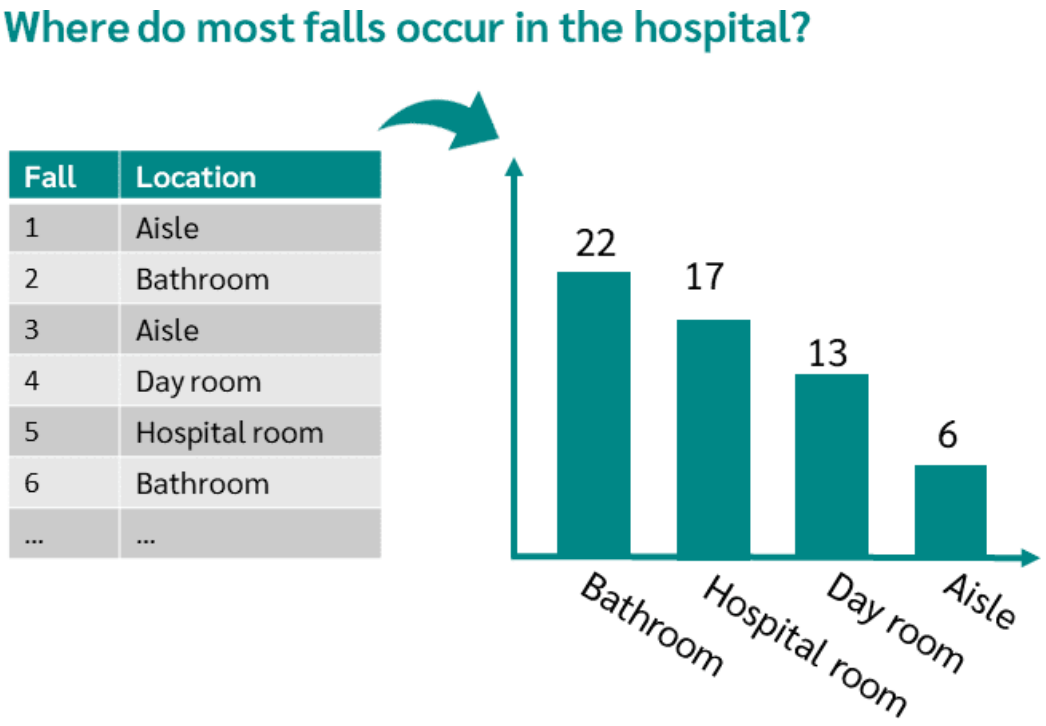


Figure 24: Bar chart falls in the hospital

8.3 Grouped bar charts

If two categorical variables are present, grouped bar charts can be created. In a grouped bar chart, the bars are grouped together. The groups result from the categories of one of the two variables, the categories of the other variable are represented by different colors.

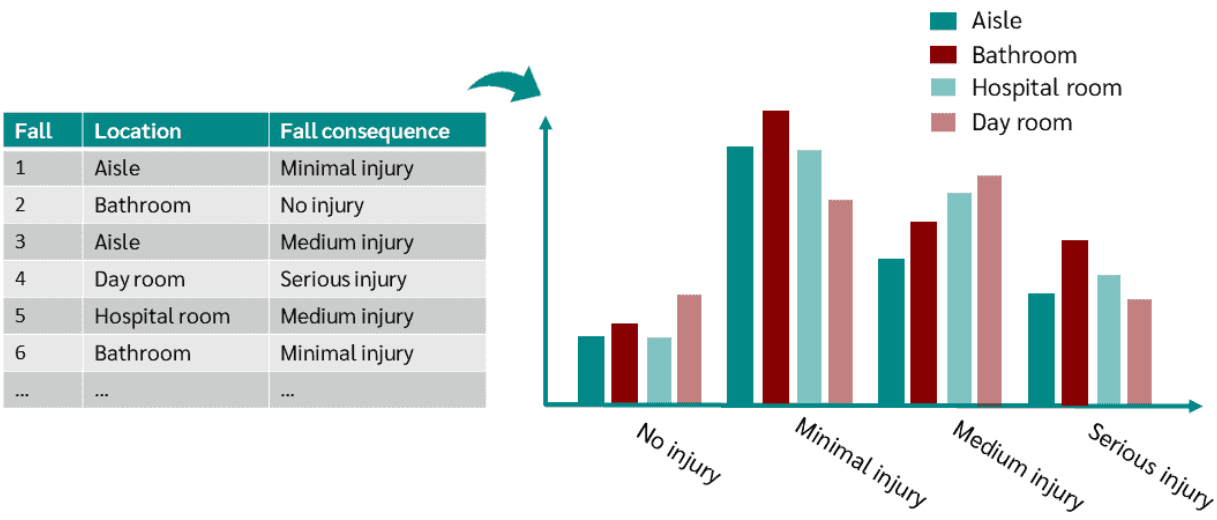


Figure 25: Grouped bar charts

In the example above, the groups are formed with the help of the variable fall sequence and the categories of the variable location are highlighted with different colors. This can of course also be reversed.

In grouped bar charts, either the frequency, the percent, or the percent in each group can be specified.

8.4 Bar chart for mean values

Of course, bar charts can be used to display not only frequencies, but also other characteristic values. In addition to frequencies, mean values are very often displayed. For this, a categorical and a metric variable must be present.

Case	Highest level of education	Monthly salary in Euro
1	College	2500
2	Bachelor's degree	3200
3	Without graduation	1800
4	Bachelor's degree	2800
5	Master's degree	3000
...	...	...

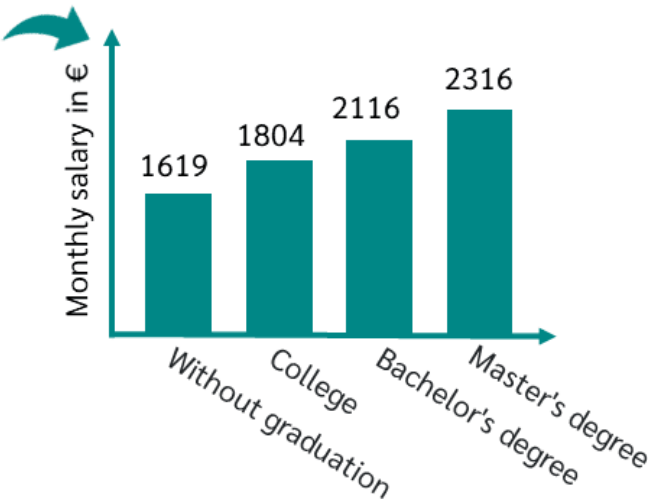


Figure 26: Bar chart for mean values

8.5 Error bar

Error bars are a graphical representation of the scatter of data. With error bars you can see how accurate your measurement is and get an overview of the range of your data!

Error bars are drawn in graphs as vertical lines above and below the measured value. The error bar is usually calculated using the standard error, standard deviation, or 95% confidence interval.

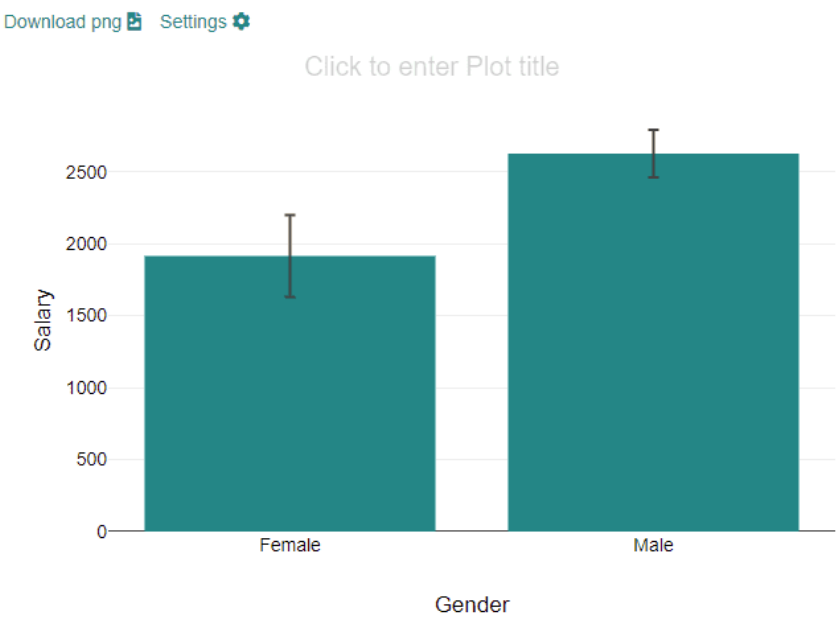


Figure 27: Error bar

8.6 Example bar chart

The procedure for creating a bar chart with DATAtab will now be explained in more detail using an example. In the example it is assumed that in a statistics course, the participants are asked which **make of car** they drive. The results can be displayed clearly and easily in a bar chart:

This is how it works in DATAtab:

To create a bar chart online, go to "Charts" and simply click on the variables you want to evaluate. The appropriate charts will then be created automatically. If you want to use your own data, just copy it into the table on DATAtab.

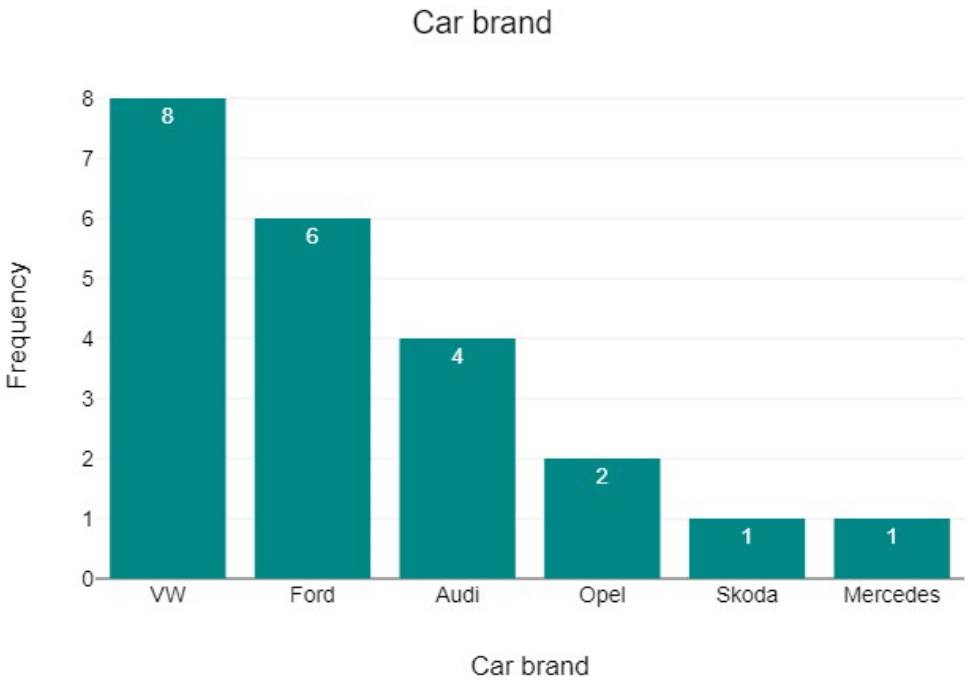


Figure 28: Example of a bar chart

If there is an **additional categorical variable**, this information can be represented with **DATAtab** by additional bars with a different coloring. For example, if **gender** is also known, the results can be displayed as follows. The blue bars referring to the "male" and the orange bars describing the "female".

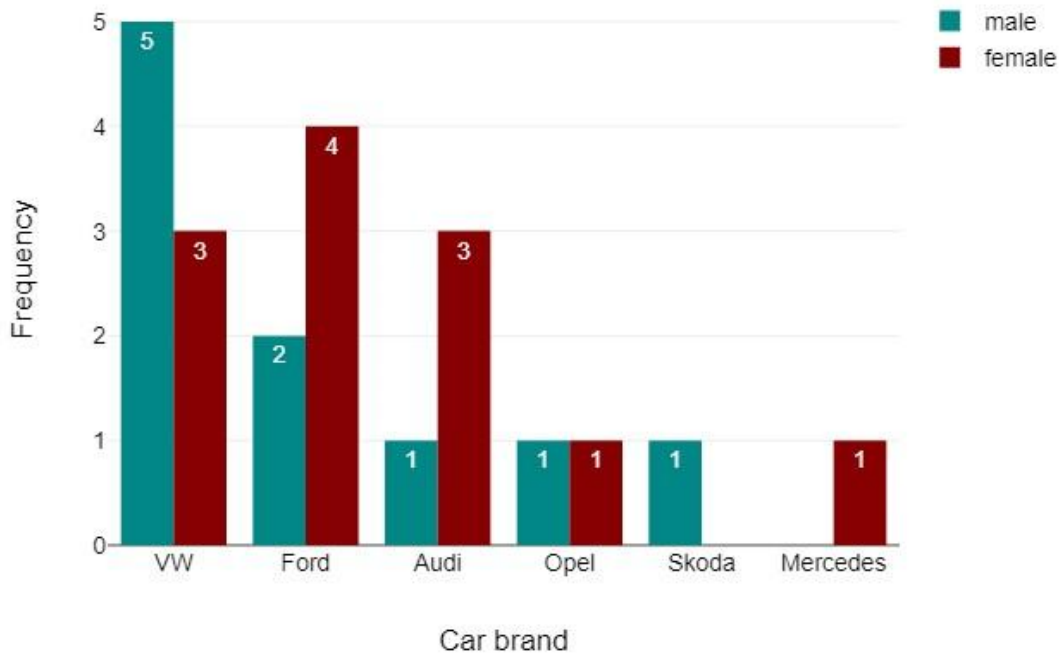


Figure 29: Nested bar chart

8.7 Histogram

A **histogram** is a graphical representation of the frequency distribution of a **metric variable.**

To display a distribution of data in a histogram, the data must first be divided into **classes**, also called **bins**. These classes or bins are then represented by **rectangles** that lie directly next to each other.

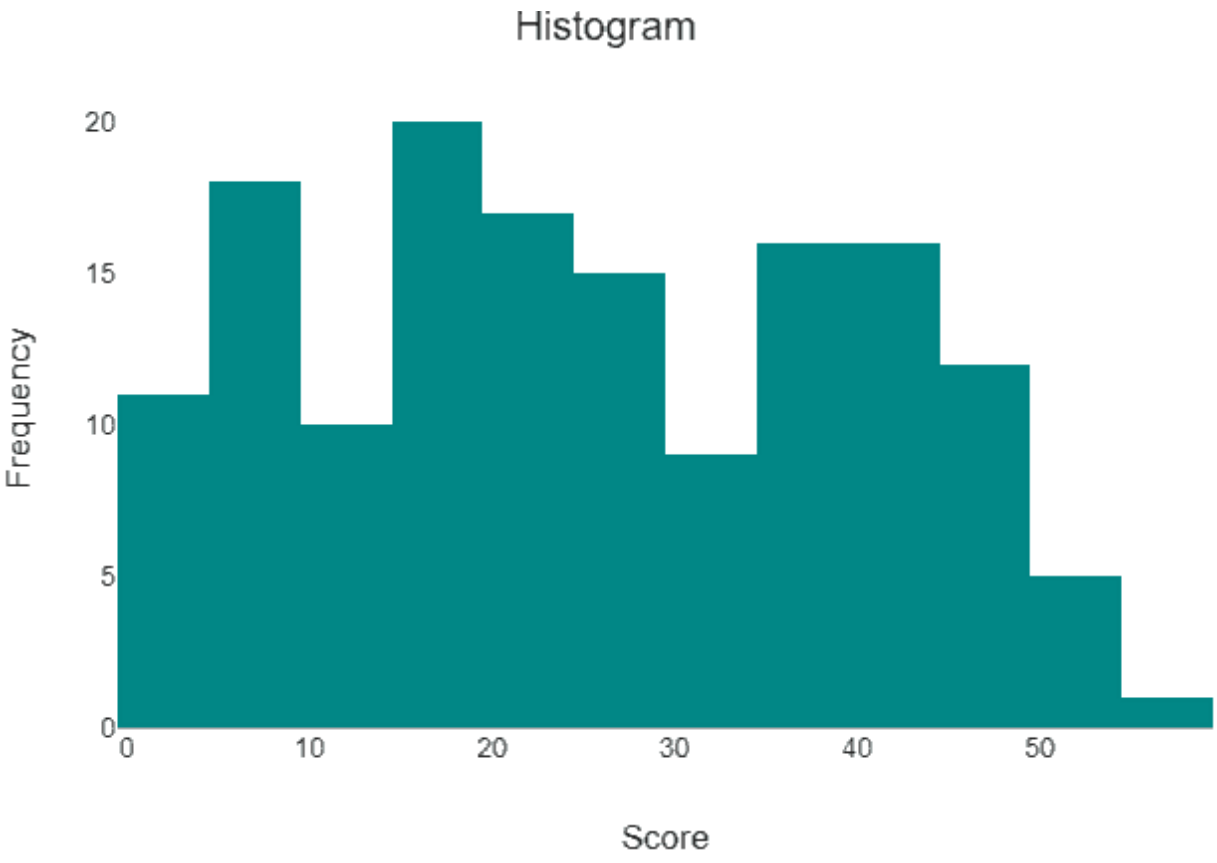


Figure 30: Example of a histogram

This is also the main difference to bar charts; in a bar chart, the data are already grouped from the outset and do not have to be divided into groups first as in a histogram. This is graphically illustrated by the fact that in a bar chart there is a space between the bars.

Accordingly, histograms are used for metric variables such as salary or age, and bar charts for ordinal or nominal variables such as gender or school grades.

8.8 Histogram example

We would like to display the frequency distribution of the results of a statistics exam graphically. For this we use a table of test scores of 12 students. You can find it here:

Student	Score
1	28
2	29
3	35
4	37
5	32
6	26
7	37
8	39
9	22
10	29
11	33
12	34

That’s how it works with DATAtab:

- Simply copy the above table into the statistics calculator, click on "Descriptive Statistics" and finally select the variable "Score".

- You will then get the following graph in DATAtab, which provides a visualization of the results of the statistics exams of 12 students.

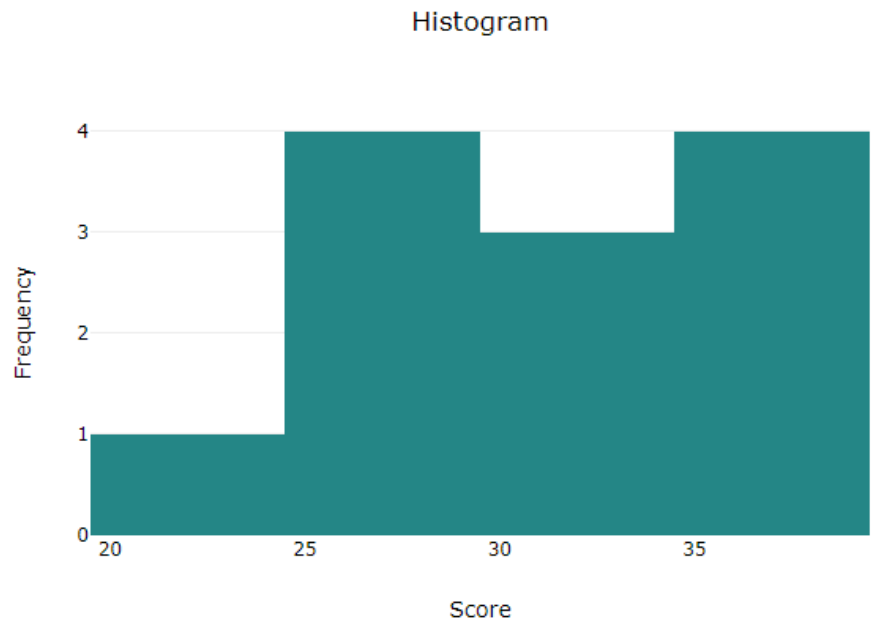


Figure 31: Example of a histogram

8.9 Bar chart vs. Histogram

A bar chart and a histogram are both types of graphical representations of data, but they are used to display different types of information.

A bar chart is used to represent discrete data, where the data is divided into separate categories. The height of each bar represents the frequency or quantity of the data that falls into that category.

A histogram, on the other hand, is used to represent continuous data, where the data is divided into a set of bins or intervals. The height of each bar represents the frequency or quantity of the data that falls into that bin or interval. The bars in a histogram are usually adjacent and there is no space between them.

In summary, the main difference between a bar chart and a histogram is the type of data they represent and the way the data is divided and displayed.

8.10 Scatter plot

Scatter plots are used in statistics to visualize correlations in data. In a scatterplot always two variables can be plotted, this is done by representing each pair of values of a case as a point in a coordinate system. If, for example, 10 persons are asked for their weight and height, the scatterplot shows 10 points.

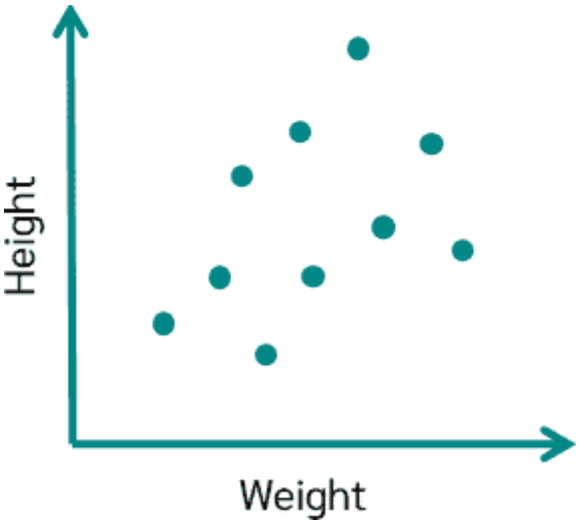


Figure 32: Example of a scatter plot

With the help of the scatterplot you get a first indication of the correlation between the two visualized variables. If high values of one variable are associated with high values of the other variable, there is a positive correlation. If high values of one variable are associated with low values of the

other variable, there is a negative correlation. If the points are randomly distributed, there is no correlation.

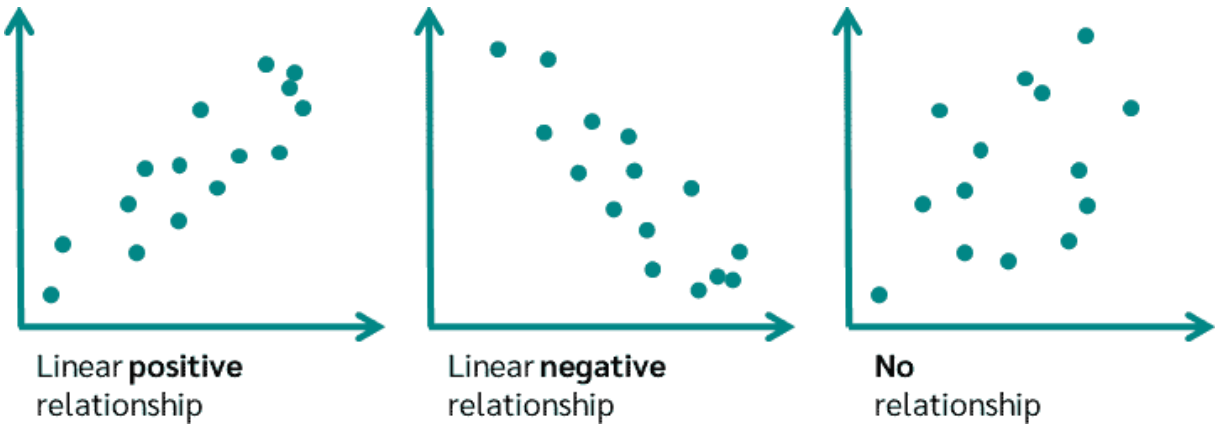


Figure 33: Interrelationships in the scatter diagram

Furthermore, there can also be a nonlinear relationship; in this case, there is a pattern in the distribution of the points, but no straight line can be drawn through the points.

8.11 Line charts

A line chart is a graph consisting of a series of data points connected by a line. It is used, for example, to show a continuous change of data over time.

In a line chart, **time or the other continuous variable** are plotted on the horizontal axis, while the values of the data to be illustrated are plotted on the vertical axis.

Line charts are particularly useful for visualizing **trends and changes over time**, and they are often used to represent economic and financial data, weather data, or scientific data.

8.12 Boxplot

What is a boxplot? With a boxplot you can graphically display a lot of information about your data. Among other things, the median, the interquartile range (IQR) and the outliers can be read in a boxplot.

The data used are mostly metric scaled, such as a person's age, annual electricity consumption, or temperature.

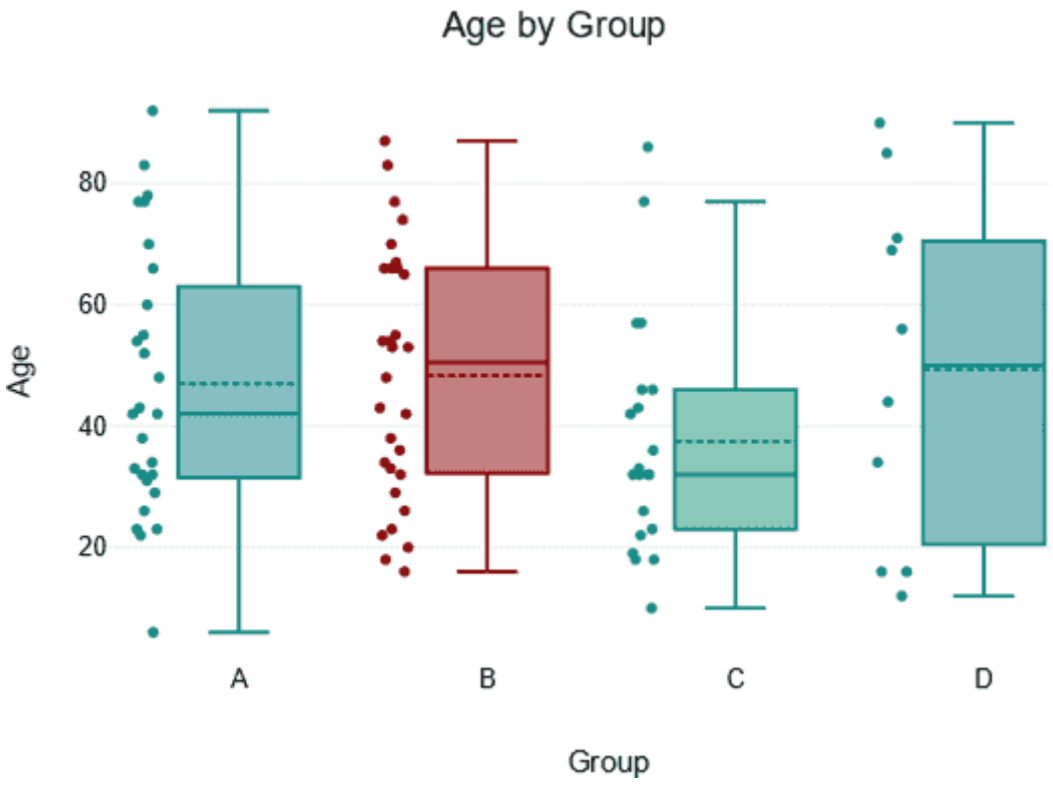
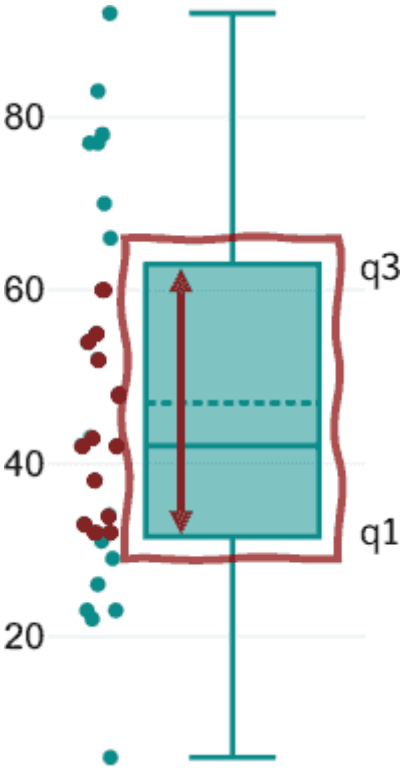


Figure 34: Boxplot example

Often a boxplot is created to compare two or more groups. For example, the salary of men and women.

Interpretation of a boxplot

The box itself indicates the range in which the middle 50% of all values lie. The lower end of the box is therefore the 1st quartile and the upper end the 3rd quartile.



Therefore below Q1 lie 25% of the data and above Q3 lie 25% of the data, in the box itself lie 50% of your data.

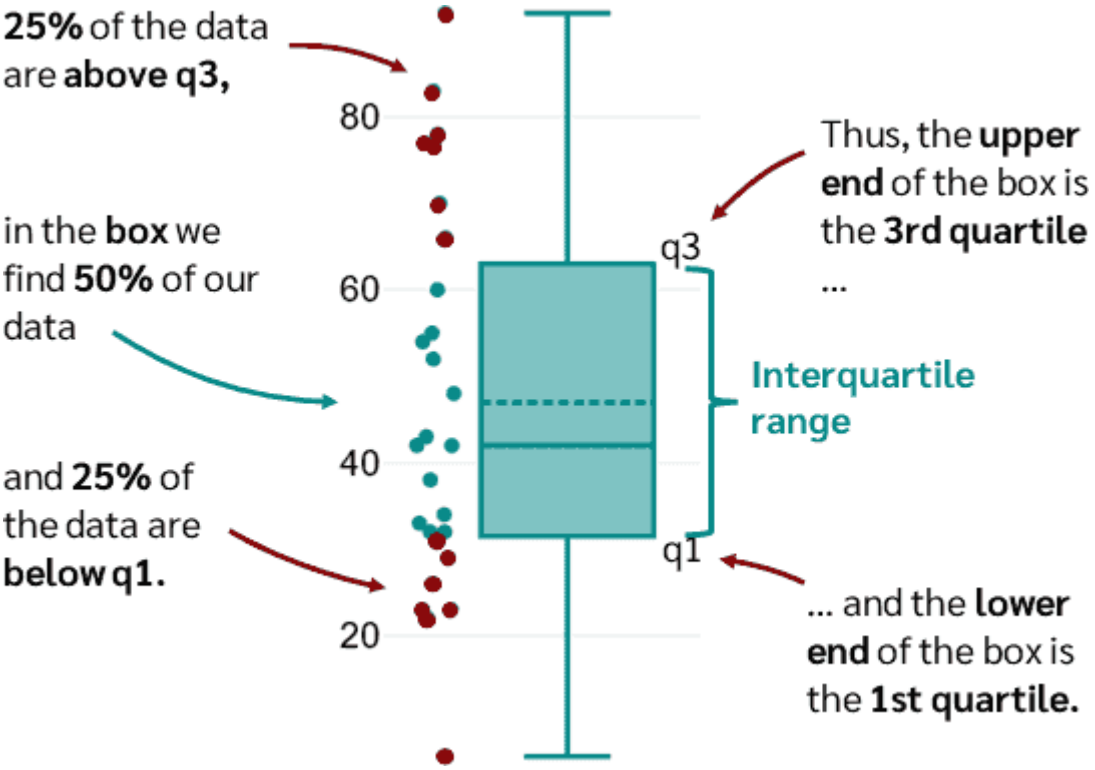

25% of the data are above q3,
in the box we find 50% of our data
and 25% of the data are below q1.
80
60
40
20
q3
q1
Thus, the upper end of the box is the 3rd quartile ...
Interquartile range
... and the lower end of the box is the 1st quartile.

Figure 35: Interpret boxplot

Let's say we look at the age of individuals in a boxplot, and Q1 is 31 years, then it means that 25% of the participants are younger than 31 years. If Q3 is 63 years, then it means that 25% of the participants are older than 63 years, 50% of the participants are therefore between 31 and 63 years old. Thus, between Q1 and Q3 is the interquartile range.

In the boxplot, the solid line indicates the median and the dashed line indicates the mean.

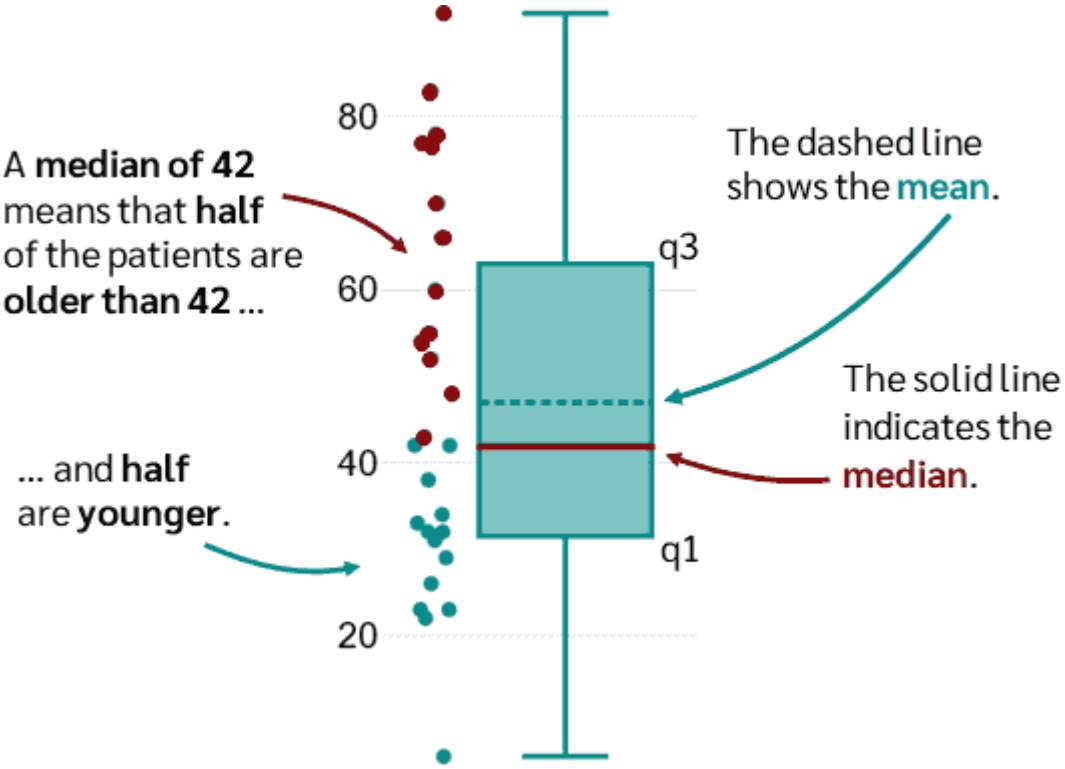


Figure 36: Interpret boxplot part 2

For example, if the median is 42, this means that half of the participants are younger than 42 and the other half are older than 42. The median thus divides the individuals into two equal groups.

The T-shaped whiskers go to the last point, which is still within 1.5 times the interquartile range. What does it mean? The T-shaped whisker is either the maximum value of your data but at most 1.5 times the interquartile range. Any observations that are more than 1.5 interquartile range (IQR) below Q1 or more than 1.5 IQR above Q3 are considered outliers. If there are no outliers, the whisker is the maximum value.

So the upper whisker is either the maximum value or 1.5 times the interquartile range. Depending on which value is smaller. The same is true for the lower whisker, which is either the minimum or 1.5 times the interquartile range.

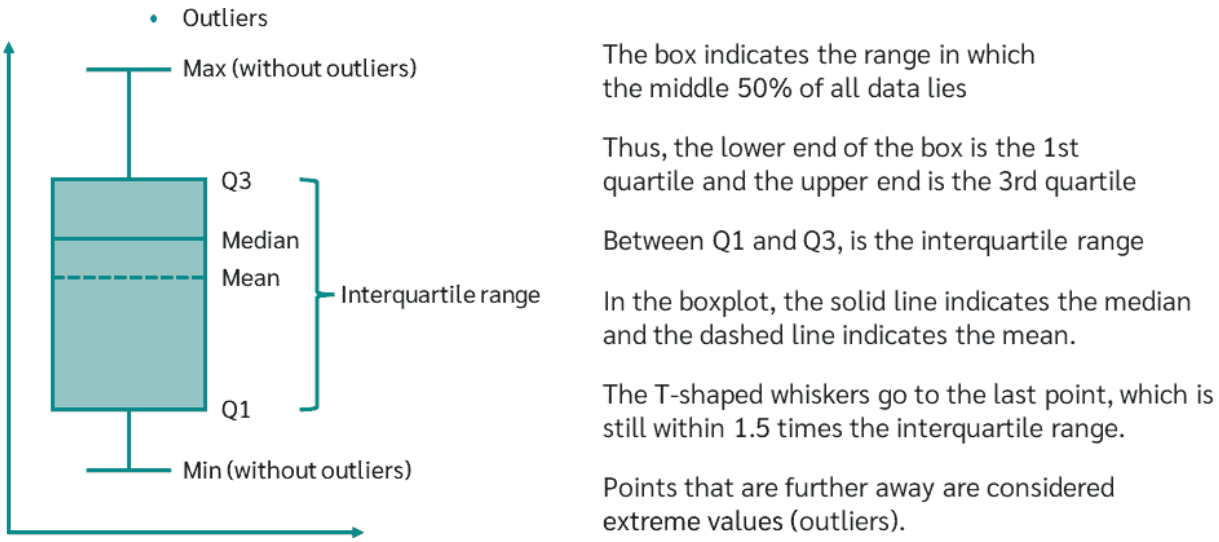

Outliers
Max (without outliers)
Q3
Median
Mean
Interquartile range
Q1
Min (without outliers)
The box indicates the range in which the middle 50% of all data lies
Thus, the lower end of the box is the 1st quartile and the upper end is the 3rd quartile
Between Q1 and Q3, is the interquartile range
In the boxplot, the solid line indicates the median and the dashed line indicates the mean.
The T-shaped whiskers go to the last point, which is still within 1.5 times the interquartile range.
Points that are further away are considered extreme values (outliers).

Figure 37: Interpret boxplot part 3

Points that are further away are considered outliers. If no point is further away than 1.5 times the interquartile range, the T-shaped whisker indicates the maximum or minimum value.

That's how it works with DATAtab:

- With DATAtab you can easily create a boxplot online.
- To do this, click on the statistics calculator,
- copy your own data into the table,
- select the tab "Descriptive" or "Charts" and
- click on the variables for which you want to create a boxplot.

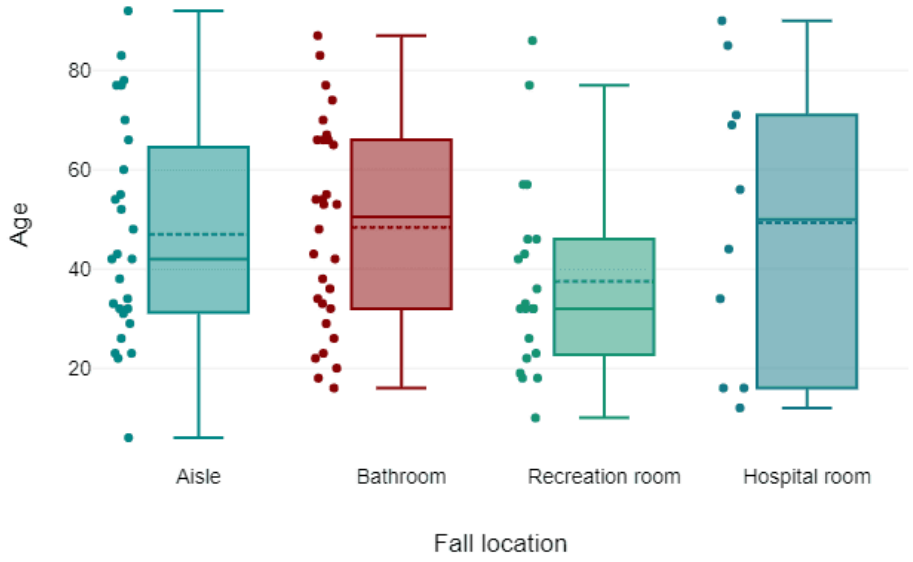


Figure 38: Interpret boxplot part 4

In the upper boxplot created with DATAtab online, the location of falls in a hospital was contrasted with the age of the persons who fell.

8.13 Bland-Altman plot

Bland-Altman plots, also known as difference plots, are a powerful graphical tool for comparing two measurement techniques and assessing the agreement between two sets of data.

The plot provides a visual representation of the difference between two measurements on the y-axis and the average of the two measurements on the x-axis.

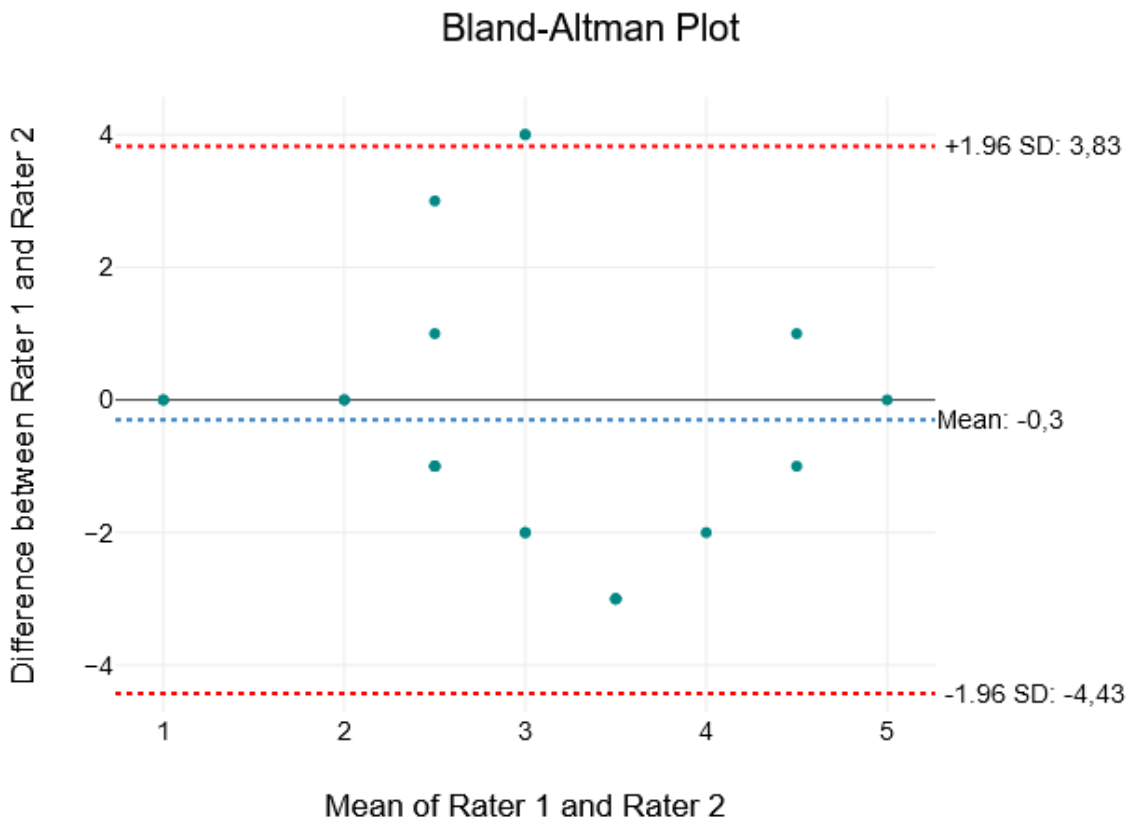


Figure 39: Bland-Altman Plot

In essence, a Bland-Altman plot is a scatter plot where the differences between two measurements are plotted against their averages. This helps to visualize the degree of agreement between the two raters and identify any systematic bias.

8.13.1 Example of Bland-Altman plot

Bland-Altman plots are widely used in medical research, industrial quality control and other fields where comparing two measurement methods is required.

In the medical field, for example, it is often necessary to compare the results of a new measurement technique with a gold standard.

The Bland-Altman plot is a powerful tool for this purpose, as it allows for the visualization of the agreement between the two methods and any systematic bias or random error.

An example of a Bland-Altman plot is to compare the measurement of blood sugar using two different measuring systems. In this case the x-axis is the mean of the two measurements and the y-axis is the difference between the two measuring systems. The plot would show the agreement or disagreement between the two measurement techniques.

8.13.2 Structure of a Bland-Altman plot

First, let's take a look at the basic structure of a Bland-Altman plot. The plot consists of a scatter plot of the differences between the two measurements against the averages of the two measurements.

A horizontal line is also included on the plot, representing the mean difference between the two measurements.

The plot also typically includes lines that represent the standard deviation, typically ±1.96 standard deviations of the differences, from the mean difference, which is used to identify any outliers in the data.

8.13.3 How can a Bland-Altman plot be used?

The Bland-Altman plot can be used to Evaluate agreement, Identify any systematic bias and Find outliers in the data.

Evaluate agreement

One of the key advantages of Bland-Altman plots is that they can be used to evaluate the agreement between two measurement techniques.

Identify any systematic bias

The plot can be used to identify any systematic bias or random error in the data. For example, if the mean difference between the two measurements is consistently positive or negative, this may indicate a systematic bias in one of the measurement techniques. Additionally, if the scatter of the points on the plot is greater than the standard deviation, this may indicate the presence of random error in the data.

8.13.4 Find outliers in the data

Another important aspect of Bland-Altman plots is that they can be used to identify outliers in the data. Outliers can have a significant impact on the results of a study, and it is important to identify them in order to understand the overall agreement between the two measurement techniques. Outliers can be identified by looking for points that fall outside of the lines representing the standard deviation from the mean difference.

That's how it works with DATAtab:

You can easily create a Bland Altman plot online with DATAtab.

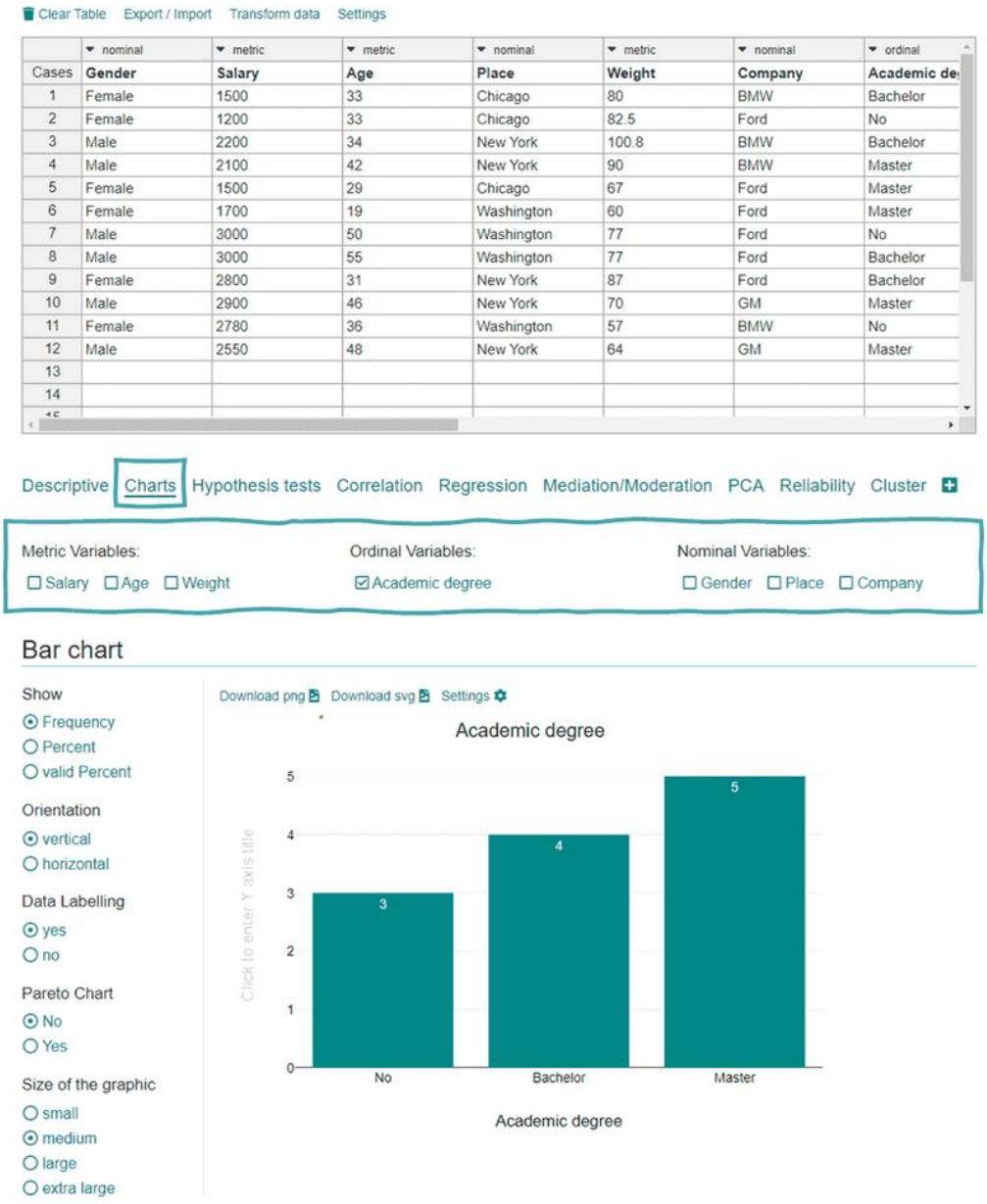
- To do this, simply copy your data into the table in the statistics calculator.
- Click on the tab Charts (create charts online) or Reliability (Reliability analysis calculator)

- Then select the desired variables for which you want to create the Bland-Altman plot online

8.14 Create charts online with DATAtab

Charts are a valuable tool for visually presenting information. With Datatab you can create your own charts online and free of charge.

To create a chart, simply select the type of chart you want to create and copy your data into the data table. The graphic below illustrates the chart creation procedure:



Cases	nominal Gender	metric Salary	metric Age	nominal Place	metric Weight	nominal Company	ordinal Academic de
1	Female	1500	33	Chicago	80	BMW	Bachelor
2	Female	1200	33	Chicago	82.5	Ford	No
3	Male	2200	34	New York	100.8	BMW	Bachelor
4	Male	2100	42	New York	90	BMW	Master
5	Female	1500	29	Chicago	67	Ford	Master
6	Female	1700	19	Washington	60	Ford	Master
7	Male	3000	50	Washington	77	Ford	No
8	Male	3000	55	Washington	77	Ford	Bachelor
9	Female	2800	31	New York	87	Ford	Bachelor
10	Male	2900	46	New York	70	GM	Master
11	Female	2780	36	Washington	57	BMW	No
12	Male	2550	48	New York	64	GM	Master
13							
14							

Figure 40: Create charts with DATAtab

To create a chart with DATAtab, simply copy your data into the table on the statistics calculator and select the variables for which you want to create a chart.

Using DATAtab you can create various charts, such as:

- Barchart
- Boxplot
- Histogram

- Scatter Plot
- Creating Violin plot
- Create Raincloud Plot
- Create Bland-Altman plot
- Create Sankey diagram
- and many more...

Which diagram you should create depends on the information you want to convey and the scale level of your data.

In statistics it is advisable to first examine the data by means of the created diagrams, this already gives an indication whether there are differences in the individual groups, for example. Subsequently, the visual results can be verified with hypothesis tests.

9. Inferential Statistics

In contrast to descriptive statistics inferential statistics want to make a statement about the population. In statistics, different types of hypotheses are distinguished and there are rules that must be observed when formulating hypotheses. These topics will now be discussed in more detail in the following chapter.

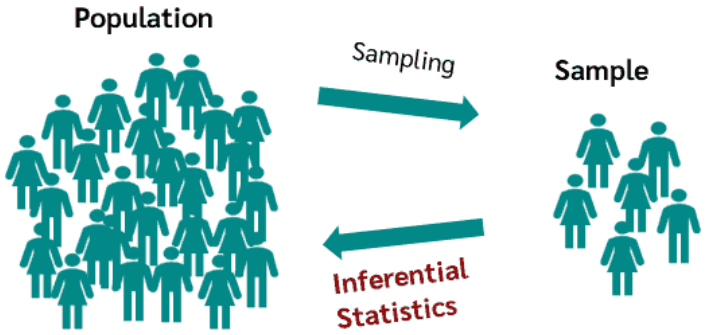


Figure 41: Aim of inferential statistics

9.1 Hypotheses

A hypothesis is an **assumption** that is neither confirmed nor disproved. In the research process, a hypothesis is made at the very beginning and the goal is to either reject or maintain the hypothesis. n order to reject or or not reject a hypothesis, data, e.g. from an experiment or a survey, are needed, which are then evaluated using a hypothesis test.

Usually, hypotheses are formulated starting from a literature review. Based on the literature review, you can then justify why you formulated the hypothesis in this way.

An example of a hypothesis is: "Men earn more than women in the same job in Germany. "

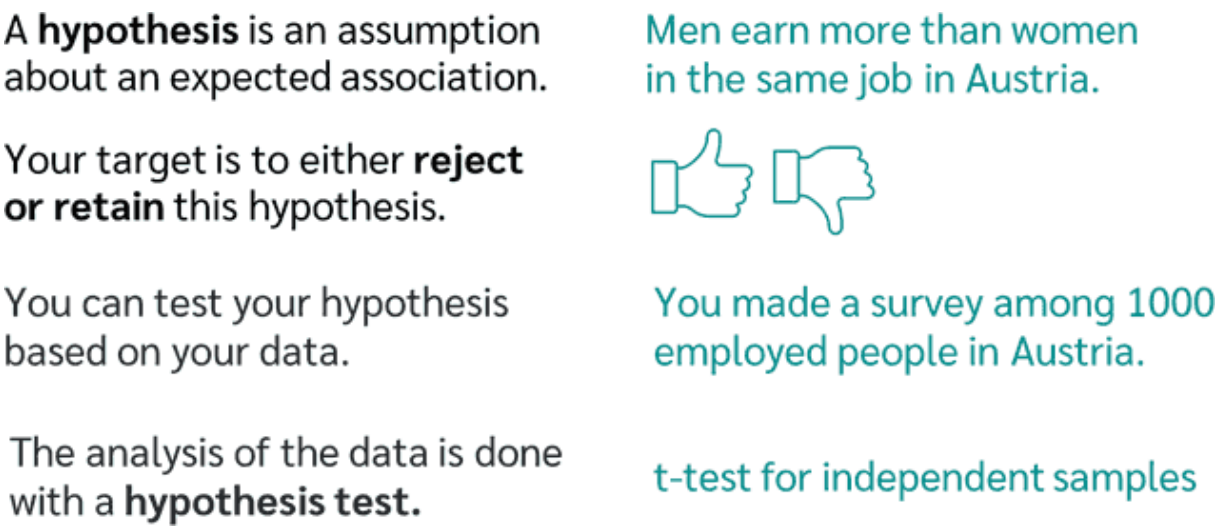

A **hypothesis** is an assumption about an expected association.
Men earn more than women in the same job in Austria.
Your target is to either **reject or retain** this hypothesis.
You can test your hypothesis based on your data.
You made a survey among 1000 employed people in Austria.
The analysis of the data is done with a **hypothesis test.**
t-test for independent samples

Figure 42: Properties of hypotheses

To test this hypothesis, you need data, e.g., survey data, and a suitable hypothesis test such as the t-test or correlation analysis. Don't worry, DATAtab will help you choose the right hypothesis test.

Formulating a hypothesis

In order to formulate a hypothesis, a **research question** must first be defined. A precisely formulated **hypothesis** about the population can then be derived from the research question, e.g., men earn more than women in the same job in Germany. Based on the hypothesis, a suitable hypothesis test is chosen to test the assumption.

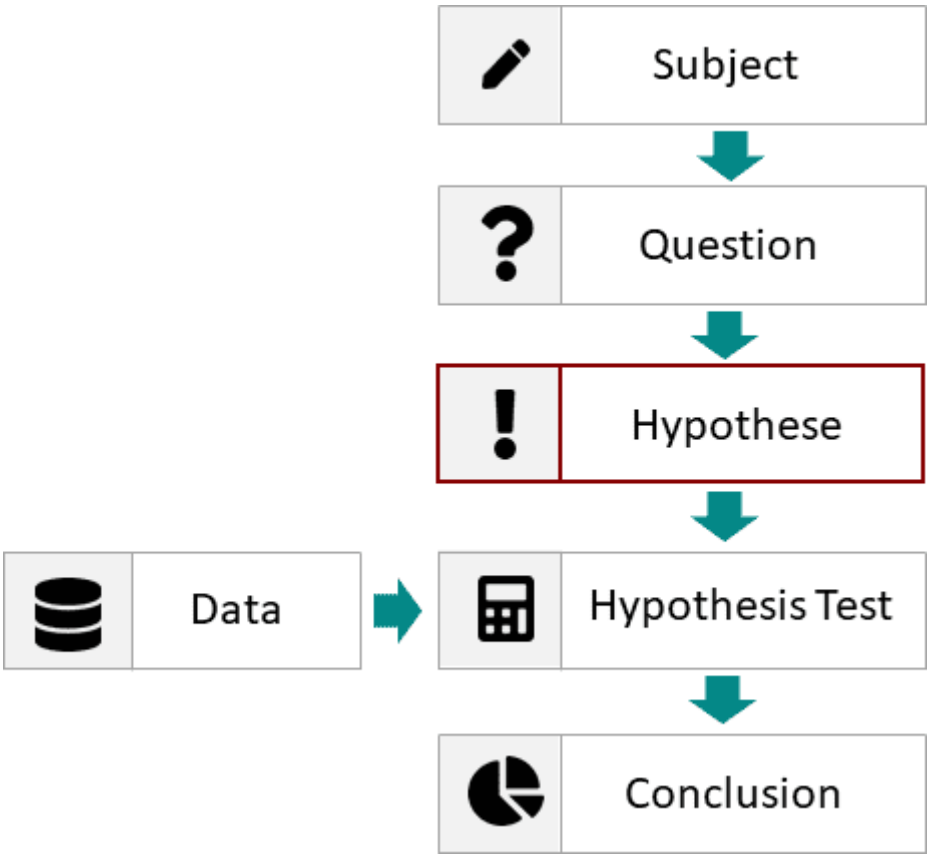


Figure 43: Hypotheses in the research process

Hypotheses are not simple statements; they are formulated in such a way that they can be tested with collected data in the course of the research process.

To test a hypothesis, it is necessary to define exactly which variables are involved and how the variables are related. Hypotheses, then, are assumptions about the cause-and-effect relationships or the associations between variables.

What is a variable?

A variable is a property of an object or event that can take different values. For example, the eye color is a variable, it is the property of the object eye and can take different values (e.g. blue or brown).

If you're doing research in the social sciences, your variables may be:

- Gender
- Income
- Attitude towards environmental protection

If you are researching in the medical field, your variables may be:

- Body weight
- Smoker status
- Heart rate

9.2 Null and alternative hypothesis

There are always two hypotheses that exactly oppose each other or assert the opposite. These opposing hypotheses are called **null** and **alternative hypothesis** and are abbreviated as H0 and H1.

The definition of the **null hypothesis** is: "The null hypothesis assumes that there is no difference between two or more groups with respect to a characteristic. "

Example:

The salary of men and women does not differ in Germany.

In contrast, the alternative hypothesis can be described as follows: "Alternative hypotheses assume that there is a difference between two or more groups. "

Example:

The salary of men and women differs in Germany.

The hypothesis that you want to test or that you have derived from the theory usually states that there is an effect e.g. gender has an effect on salary. This hypothesis is called an alternative hypothesis.

The null hypothesis usually states that there is no effect e.g. gender has no effect on salary. In a hypothesis test, only the null hypothesis can be tested; the goal is to find out whether the null hypothesis is rejected or not.

9.3 Difference and correlation hypotheses

What types of hypotheses are there? The most common distinction is between **difference and correlation hypotheses** as well as **directional and non-directional hypotheses**.

Difference hypotheses are used when different groups are to be distinguished, e.g., the group of men and the group of women. Correlation hypotheses are used when the relationship or correlation between variables is to be tested, e.g., the relationship between age and height.

Difference hypotheses

Difference hypotheses test whether there is a difference between two or more groups.

Figure 44: Difference hypotheses

Examples of difference hypotheses:

- The "group" of men earns more than the "group" of women.
- Smokers have a higher risk of heart attack than nonsmokers.
- There is a difference between Germany, Austria, and France in terms of hours worked per week.

Thus, one variable is always a **categorical variable, e.g.,** gender (male, female), smoking status (smoker, nonsmoker), or country (Germany, Austria,

and France); the other variable is at least **ordinally scaled**, e.g., salary, percent risk of heart attack, or hours worked per week.

Correlation hypotheses

Correlation hypotheses test correlations between at least two variables.

Figure 45: Correlation hypotheses

Here you can find some examples:

- The taller a person is, the heavier he or she is.
- The more horsepower a car has, the higher its fuel consumption.
- The better the math grade, the higher the future salary.

As can be seen from the examples, correlation hypotheses often take the form "The more..., the higher/lower.... ". Thus, at least two ordinally scaled variables are being examined.

9.4 Directional and undirectional hypotheses

Hypotheses are divided into **directional and undirectional (non-directional)** or **one-sided and two-sided hypotheses.** If words such as "better than" or "worse than" occur in the hypothesis, the hypothesis is usually directed.

Figure 46: Directional and undirectional hypotheses

In the case of an undirectional hypothesis, one often finds building blocks such as "there is a difference between" in the formulation, but it is not stated in which direction the difference lies.

- With an **undirectional hypothesis**, the only thing of interest is whether there is a difference in a value between the groups under consideration.
- In a **directional hypothesis**, what is of interest is whether one group has a higher or lower value than the other.

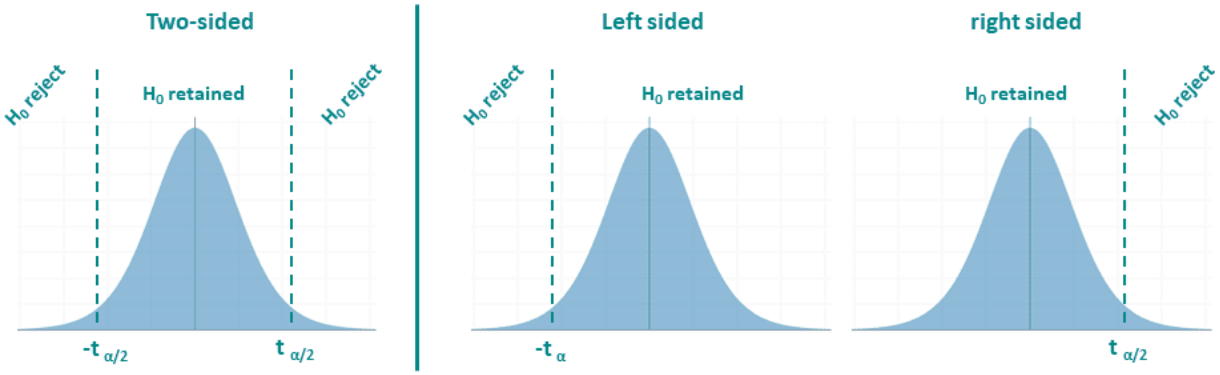


Figure 47: One-sided and two-sided testing

Undirectional hypotheses

Undirectional hypotheses test whether there is a relationship or a difference, and it does not matter in which direction the correlation or difference goes. In the case of a difference hypothesis, this means there is a difference between two groups, but it does not say whether one of the groups has a higher value.

- There is a difference between the salary of men and women (but it is not said who earns more!).
- The risk of heart attack differs between smokers and non-smokers (but it is not said who has the higher risk!).

In terms of a correlation hypothesis, this means there is a relationship or correlation between two variables, but it does not say whether this relationship is positive or negative.

- There is a correlation between height and weight.
- There is a correlation between horsepower and fuel consumption in cars.

In both cases it is not said whether this correlation is positive or negative!

Directional hypotheses

Directional hypotheses additionally indicate the direction of the correlation or the difference. In the case of the **difference hypothesis** a statement is made as to which group has a higher or lower value.

- Men earn **more than** women.
- Smokers have a **higher** risk of heart attack than nonsmokers.

A correlation hypothesis indicates whether the correlation is positive or negative.

- **The taller** a person is, the **heavier** he is.
- **The more** horsepower a car has, the **higher** its fuel consumption.

A one-sided or directional alternative hypothesis includes only values that differ in one direction from the value of the null hypothesis.

The p-value for directional hypotheses

Usually, statistical software always calculates the non-directional test and then also outputs the p-value for this.

To obtain the p-value for the directional hypothesis, it must first be checked whether the effect is in the right direction. Then the p-value must be divided by two. This is because the significance level is not split on two sides, but only on one side. More about this in the tutorial about the p-value.

If you select "one-tailed" in DATAtab for the calculated hypothesis test, the conversion is done automatically and you only need to read the result.

9.5 Hypothesis Testing

Now that the basic rules of hypothesis formulation and the different types of hypotheses have been discussed, the next section deals with the goals of hypotheses, the basics of hypothesis testing, and the logic of statistical inference.

The starting position of hypothesis testing is that in your thesis you have formulated several hypotheses that you want to test in order to make a statement about the population.

A hypothesis test is used whenever you want to test a hypothesis about the population with the help of a sample. So, whenever you want to prove or say something about the population with a sample, hypothesis tests are used.

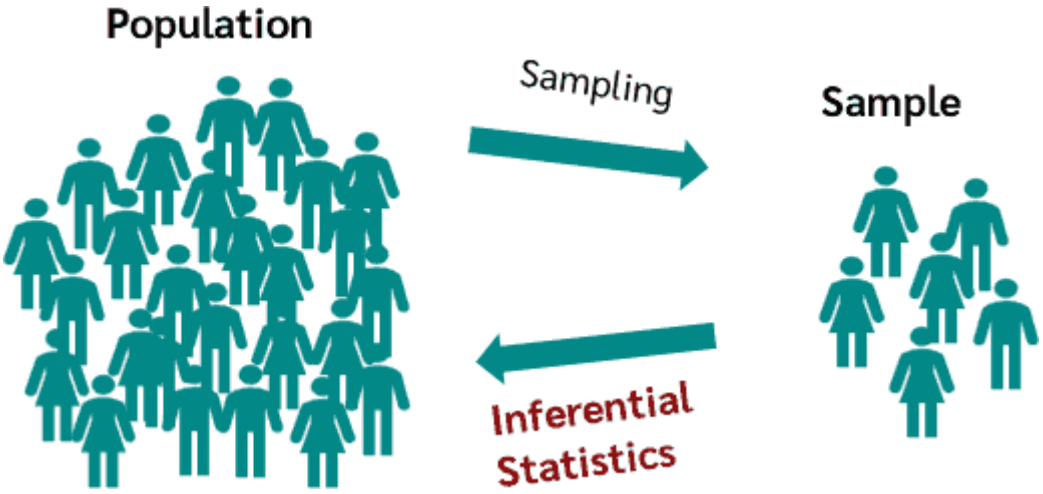


Figure 48: Logic of statistical inference

A possible example is that the company My-Muesli would like to know whether their produced muesli bars really weigh 250g. For this purpose, a random sample is taken, and a hypothesis test is then used to draw conclusions about all the muesli bars produced.

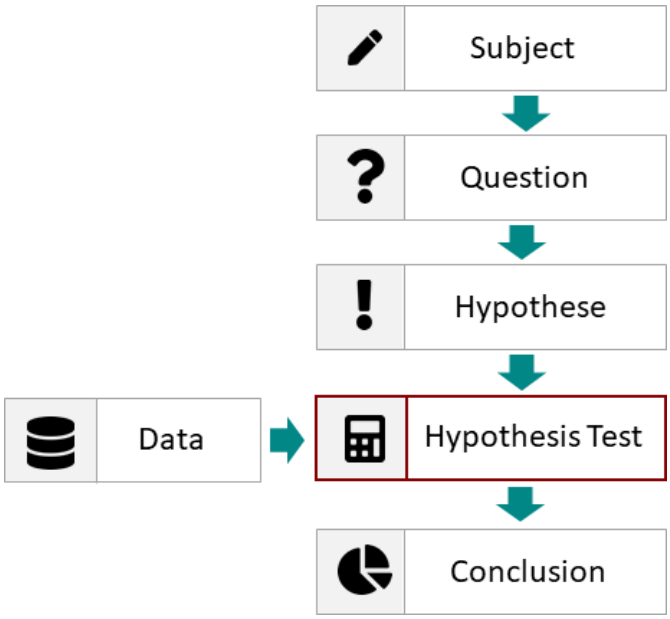


Figure 49: Hypothesis testing and research process

9.5.1 Hypothesis testing and the null hypothesis

As we know from the previous tutorial on hypotheses, there is always a null and an alternative hypothesis. In "classical" inferential statistics, the null hypothesis is always tested using a hypothesis test. The hypothesis is tested to see if there is no difference or no relationship.

If you want to be 100% accurate (what is not always the case in practice), the null hypothesis H0 can only ever be “rejected” or “not rejected” using a hypothesis test. The non-rejection of H0 is not a sufficient reason to conclude that H0 is true. Thus, it must always be "H0 was not rejected" and not "H0 was retained."

9.5.2 The uncertainty in hypothesis testing

Whether an assumption or hypothesis about the population is accepted or rejected by hypothesis testing can only ever be determined with a certain **probability of error.** But why does the probability of error exist at all?

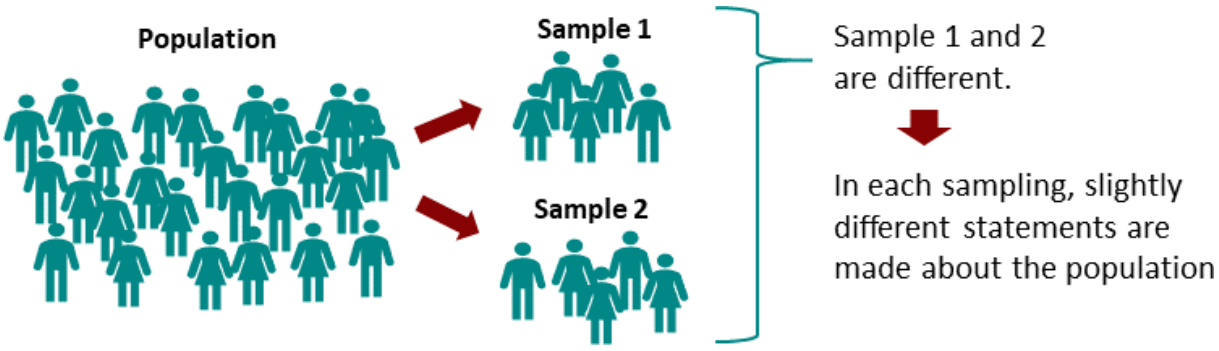

Population
Sample 1
Sample 2
Sample 1 and 2 are different.
In each sampling, slightly different statements are made about the population

Figure 50: Uncertainty in hypothesis testing

Here is the short answer: Each time you take a sample, of course, you get a different sample, which means that the results are different every time. In the worst case, a sample is taken that happens to deviate very strongly from the population and the wrong statement is made. Therefore there is always a probability of error for every statement or hypothesis.

9.5.3 Level of significance or probability of error

A hypothesis test can never reject the null hypothesis with absolute certainty. There is always a certain probability of error that the null hypothesis is rejected even though it is actually true. This **probability of error** is called the **significance level or α.**

The significance level is used to decide whether the null hypothesis should be rejected or not. If the p-value is smaller than the significance level, the null hypothesis is to be rejected; otherwise, it is not to be rejected.

Usually, a significance level of **5% or 1%** is set. If a significance level of 5% is set, it means that it is 5% likely to reject the null hypothesis even though it is actually true.

Illustrated by the two-sample t-test, this means: The observed means of two samples have a certain distance to each other. The greater the observed distance between the mean values, the less likely it is that both samples come from the same population. The question now is, at what point is it "unlikely enough" to reject the null hypothesis? If a significance level of 5% is set, at 5% it is "unlikely enough" to reject the null hypothesis.

The probability that two samples are drawn from a population and that they have the observed mean difference, or even a greater one, is indicated by the p-value. Accordingly, if the p-value is less than the significance level, the null hypothesis is rejected; if the p-value is greater than the significance level, the null hypothesis is not rejected.

If, for example, a p-value of 0.04 results, the probability that two groups with an observed mean distance or an even greater distance come from the same population is 4%. The p-value is thus smaller than the significance level of 5% and thus the null hypothesis is rejected.

It is important that the significance level is always determined prior to the investigation and may not be changed subsequently in order to obtain the "desired" statement after all. To ensure a certain degree of comparability, the significance level is usually 5% or 1%.

- **$\alpha \leq 1\%$ highly significant (h.s.)**
- **$\alpha \leq 5\%$ significant (s.)**
- **$\alpha > 5\%$ not significant (n.s.)**

9.5.4 Example significance level and p-value

The significance of significance level and p-value will now be illustrated by an example.

H0: Men and women in Austria do not differ in their average monthly net income.

To test this hypothesis, a significance level of 5% is set and a survey is conducted asking 600 women and 600 men about their monthly net income. An independent t-test gives a p-value of 0.04

The p-value 0.04 is less than the significance level of 0.05, thus we rejecting the null hypothesis. Based on the data collected, we have sufficient evidence that there is a statistically significant difference in average monthly next income for the population of men and women in Austria.

9.5.5 Types of errors

Because a hypothesis can only be rejected with a certain probability, different types of errors occur. Due to the sample selection, it can happen that the null hypothesis is rejected by chance, although in reality there is no difference, i.e. the null hypothesis is valid. Conversely, the result of the hypothesis test can also be that the null hypothesis is not rejected, although in reality there is a difference and thus the alternative hypothesis is actually true.

Accordingly, two types of errors arise in hypothesis testing:

- **Type 1 error:** If the alternative hypothesis is accepted although the null hypothesis is valid.
- **Type 2 error:** If the null hypothesis is retained although the alternative hypothesis applies.

Overall, the following cases arise:

	Decision	
	for H_0	against H_0
H_0 true	Right	Type 1 error
H_0 false	Type 2 error	Right

Figure 51: Types of errors in hypothesis tests

9.5.6 Significance vs effect size

We now know that we usually accept the alternative hypothesis when the p-value is less than 0.05. We then assume that there is an effect, e.g., a difference between two groups.

However, it is important to keep in mind that just because an effect is **statistically significant** does not mean that the **effect is relevant**.

If a very large sample is taken and the sample has a very small spread, even a very small difference between two groups may be significant, but it may not be relevant to you.

Example

A company sells frozen pizza and wants to test whether higher quality packaging leads to increased sales.

Based on the data collected, it shows that the p-value is less than 0.05 and therefore there is a statistically significant increase.

So the company can assume that the higher quality packaging will increase the sales statistically significant. It is less than 5% probable that this increase or an even greater increase would occur if the packaging had no influence.

But now the question is whether the increase is also economically relevant. It may be that the income from the increased sales figures does not compensate for the higher costs of the packaging.

Therefore, one should always consider both whether an effect is significant and whether the effect is relevant at all.

9.5.7 Choosing the appropriate hypothesis test

In order to test hypotheses, various test procedures or hypothesis tests are available. In the course of selecting the appropriate test procedure, two further criteria are decisive.

On the one hand, these are subdivided according to the **scale level of** the characteristics of interest:

- Nominal
- Ordinal
- Interval

and on the other hand, according to how many **samples or groups** are available and how the samples are related to each other.

That's how it works with DATAtab:

- DATAtab helps you find the right test; you just need to select the data you want to evaluate.
- Depending on the scale level of your data, DATAtab will suggest the appropriate test.

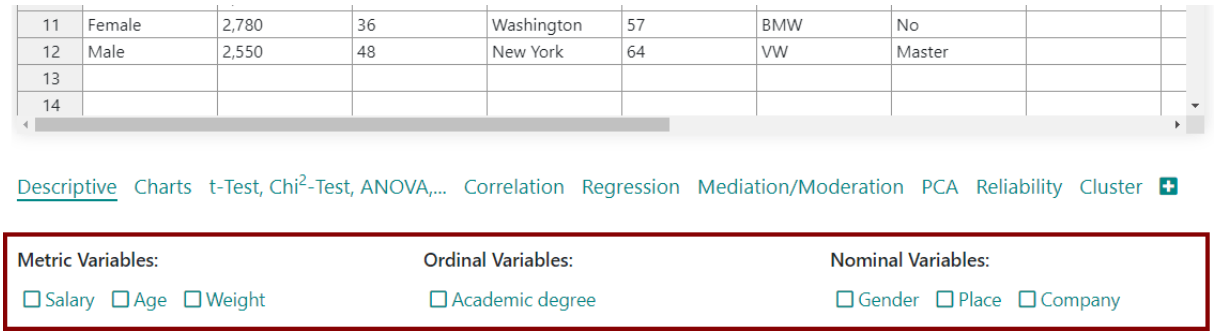

11 Female 2,780 36 Washington 57 BMW No
12 Male 2,550 48 New York 64 VW Master
13
14
Descriptive Charts t-Test, Chi²-Test, ANOVA,... Correlation Regression Mediation/Moderation PCA Reliability Cluster
Metric Variables:
Salary Age Weight
Ordinal Variables:
Academic degree
Nominal Variables:
Gender Place Company

Figure 52: Selection of the scale level with DATAtab

Depending on which variables are selected, one of the following statistical tests will be calculated:

- t-test one sample
- t-test independent samples
- t-test dependent samples
- Chi Square-Test
- Binomial test
- ANOVA with/without rep. measures
- 2 way ANOVA with/without rep. measures
- Wilcoxon-Test
- Mann-Whitney U-Test
- Friedman Test
- Kruskal-Wallis Test

The following table lists the relevant test procedures. If you know the scale level of the variables in your hypothesis, you can see in the table which test could fit!

	Level of measurement		
	nominal	**ordinal**	**metric**
Binomial test	1 x nominal		
t-test for one sample			1 x metric
Chi-Square Test	1 x or 2 x nominal		
t-test for independent samples	1 x nominal with two categories		1 x metric
Mann-Whitney U test	1 x nominal with two categories	1 x ordinal	
One-factor analysis of variance	1 x nominal with more than two categories		1 x metric
Kruskal-Wallis-Test	1 x nominal with more than two categories	1 x ordinal	
Pearson correlation			2 x metric
Spearman correlation		2 x ordinal	
Point biserial correlation	1 x nominal with two categories		1 x metric
t-test for dependent samples			2 x metric
Wilcoxon-Test		2 x ordinal	
Analysis of variance for repeated measurements			more than 2 x metric
Friedman Test		more than 2 x ordinal	

Figure 53: Overview Hypothesis tests

9.5.8 Examples for hypothesis tests

Having discussed the selection criteria for the appropriate hypothesis test, here are some examples of questions that can be answered using the appropriate test procedure.

Independet sample t-test

Within the framework of a t-test for independent samples, the following question, for example, can be analyzed:

"Is there a difference in the average number of burglaries (dependent variable) in homes with and without alarms (independent variable with 2 groups)? "

Paired t-test

A t-test for dependent samples could focus on the following aspect: "Does the consumption of cigarettes have a negative effect on the blood pressure of students? " In this case, the blood pressure of smoking students is measured and finally again the blood pressure after they have quit smoking. (before-after measurement)

ANOVA

Finally, single-factor analyses of variance are often calculated, which might ask, for example, "Do people who live in small, medium, or large cities (independent variable with three groups) differ with respect to their health consciousness (dependent variable)?"

9.6 The p-value

In the following chapter, we will first discuss the p-value, an important characteristic of statistical testing, and illustrate its calculation with examples. This is followed by the topics of significance level and distribution functions to illustrate the logic of statistical inference.

9.6.1 Defining the p-value

The p-value indicates the probability that the observed result or an even more extreme result will occur if the null hypothesis is true.

The p-value is used to decide whether the null hypothesis is rejected or retained (not rejected). If the p-value is smaller than the defined significance level (often 5%), the null hypothesis is rejected, otherwise not.

You want to make a statement about the population and have set up a hypothesis for this. Since it is usually not possible to survey the entire population, you survey a sample. Now this sample, due to chance, will most likely deviate from the population.

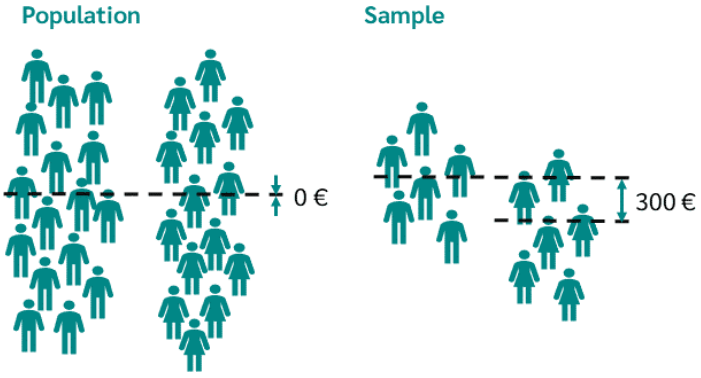


Figure 54: Example p-value

If the null hypothesis applies in your population, e.g. the salary of men and women does not differ in Germany, then there will certainly still be a difference in the sample, e.g. a difference of 300 euros per month. Now the p-value tells you how likely it is that a difference of 300 euros or more will occur by chance in the sample if there is no difference in the population.

If the result is a very small probability, you can of course ask yourself whether the assumption about the population is true at all.

If the p-value is 3%, for example, then it is only 3% likely that a sample is drawn in which the salaries of men and women differ by more than 300 euros.

9.6.2 Using the p-value

The p-value is used to either reject or retain (not reject) the null hypothesis in a hypothesis test. If the calculated p-value is smaller than the significance level, which in most cases is 5%, then the null hypothesis is rejected, otherwise it is retained.

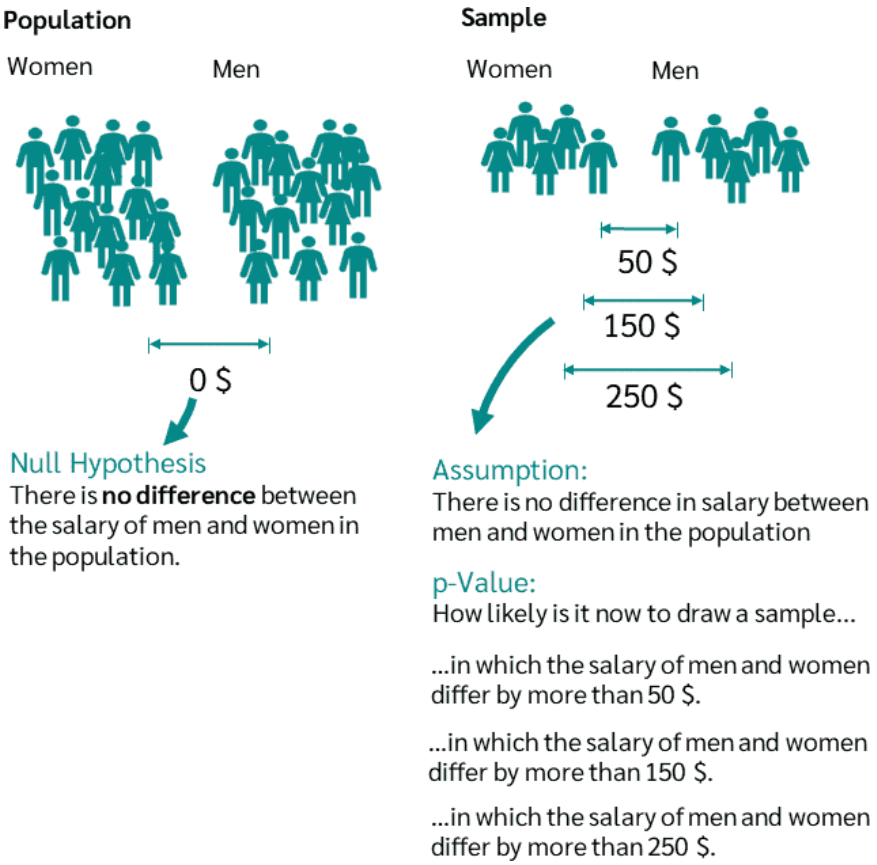

Population
Women
Men
0 $
Null Hypothesis
There is no difference between the salary of men and women in the population.
Sample
Women
Men
50 $
150 $
250 $
Assumption:
There is no difference in salary between men and women in the population
p-Value:
How likely is it now to draw a sample...
...in which the salary of men and women differ by more than 50 $.
...in which the salary of men and women differ by more than 150 $.
...in which the salary of men and women differ by more than 250 $.

Figure 55: The interpretation of the p-value

In the following section, we will now give you an example of the role of the p-value:

- The null hypothesis is that there is no difference between the salary of men and women.
- Now a sample is drawn with the salary of men and women. These are our observed results.
- We assume that the null hypothesis is true, that is, that there is no difference between the salary of men and women.
- In the observed result (sample), it has now emerged that men earn 150€ more per month than women.
- The p-value now indicates how likely it is to draw a sample in which the salary of men and women differs by 150€ or more, although there is no difference in the population.
- If the p-value is 0.04, it is only 4% likely to draw a sample that is €150 or more apart if there is no difference in salary in the population.

Let's say, in the case above we have a p-value of 0.04 or 4%, what does this p-value mean now? Put simply, the result means that if there is no difference in salary in the population, it is only 4% likely to draw a sample that is 150€ or more apart.

The probability of 4% is of course very low, so that one can ask whether it is at all true that men and women earn the same amount in the population, or whether this hypothesis should not rather be discarded.

The question from when the null hypothesis is discarded answers the significance level.

9.6.3 Significance level

The significance level is determined before the test. If the calculated p-value is below this value, the null hypothesis is rejected, otherwise it is retained. As a rule, a **significance level of 5 %** is chosen.

- **alpha < 0,01: very significant result**
- **alpha < 0.05: significant result**
- **alpha > 0,05: not significant result**

The significance level thus indicates the probability of a 1st type error. What does this mean? If there is a p-value of 5% and the null hypothesis is rejected, the probability that the null hypothesis is valid is 5%, i.e. there is a 5% probability of making a mistake.

If the critical value is reduced to 1%, the probability of error is accordingly only 1%, but it is also more difficult to confirm the alternative hypothesis.

9.6.4 One-tailed p-values

Let's say you are examining the reaction time of two groups. Then it is often not of interest whether there is a difference between the two groups, but whether one group has a larger or smaller value than the other. In this case, you would have a directional hypothesis and then calculate what is called a one-sided p-value.

A **one-tailed p-value** includes values more extreme than the obtained result in one direction, that direction having been stated in advance.

A **two-tailed p-value** includes values more extreme in both positive and negative directions.

The one-sided p-value is then obtained by dividing the two-sided p-value by 2. Here, of course, care must be taken whether the difference or effect under consideration is at all in the direction of the alternative hypothesis.

Example

Your alternative hypothesis is that group A has greater reaction time values than group B. When analyzing your data you get a two-sided p-value of 0.04.

Now you have to check whether group A really has larger values in your data. If this is the case, the two-sided p-value is divided by two, so you get 0.02.

If this is not the case and the effect or difference goes exactly in the other direction than formulated in the alternative hypothesis, your p-value is 1-0.02, i.e. 0.98.

Don't worry, if you use DATAtab, you can specify what kind of hypothesis you have, and DATAtab will help you evaluate it.

9.6.5 Calculate p-value

To calculate the p-value, a suitable hypothesis test must first be found. Once the appropriate hypothesis test is found, you can calculate the p-value in the statistics calculator on DATAtab.

The best-known hypothesis tests are:

- t-test
- correlation analysis
- chi-square test
- analysis of variance

For the **calculation of the p-value,** a distribution function is required. If this distribution function is known, it can be determined how likely it is that a drawn sample is less than or equal to a considered value. Classical representatives of these distributions are the t-distribution, the chi-square distribution, the z-distrubition and the F-distribution. These t-distribution and chi-square distribution are shown below.

The **t-distribution** is a test distribution that is used to calculate t-test statistics. If you want to test a hypothesis with the t-test, the t-value from the test result is compared with the critical t-value.

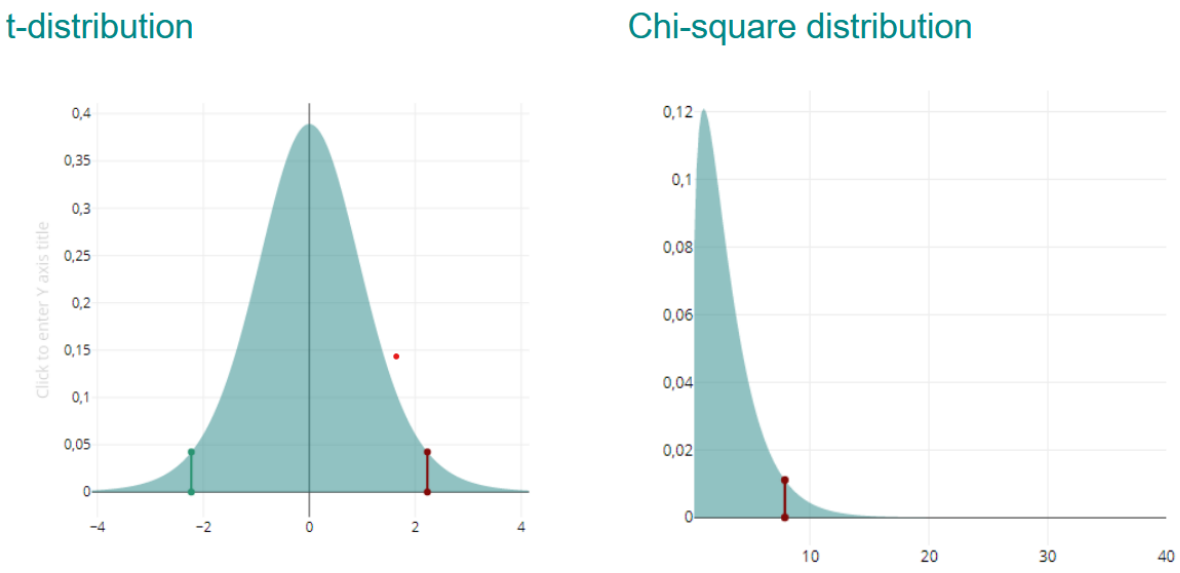


Figure 56: The t-distribution and chi-square distribution

The **chi-square distribution** is needed to analyze frequencies. With the help of the chi-square distribution, a critical chi-square value can be found. The critical value indicates the point at which a deviation is unlikely, assuming that the null hypothesis holds.

9.6.6 Statistical tests and the p-value

In order to reject or maintain a hypothesis one needs the p-value. The procedure how the p-value is used for statistical tests, is now the following:

- Definition of the critical p-value or the significance level e.g. 5 %
- Definition of a statistical test procedure e.g. t-tests or Correlation analysis
- Calculation of the test statistics from the sample e.g. the t value in the t-test
- Determination of the p-value for the test statistics e.g. p-value for given t in t-test
- Check whether the p-value is above or below the specified critical p-value e.g. p-value 1% leaks below the critical value of 5%

9.6.7 Specify the p-value

It is not always clear how the p-value from a hypothesis test should be presented in scientific papers. Different statistical programs output the p-value in different formats, some of which are not needed for writing.

In theory, the p-value could be reduced to 'statistically significant' or 'not statistically significant', which may be appropriate if the estimate and a 95% confidence interval are also provided, but this is generally not enough information.

For example, if two p-values are close to the 0.05 level, one just above and one just below, the interpretation of the two values should not be very different. If we reduce these p-values to a binary response and say that one is significant and the other is not, without qualifying that statement, we risk misrepresenting the evidence provided by the tests.

It is therefore advisable to report the actual p-value, as this gives the reader the best possible insight into the results.

It is common to use asterisks to indicate statistical significance: * for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$. In general, the following is recommended for p-values:

- Give the actual p-value if possible
- Rounding: Two significant digits are usually sufficient

10. Checking assumptions of statistical tests

Statistical test procedures usually have some assumptions that should be fulfilled in order for the test to be valid. You should check these requirements at the beginning of your analyses. We can test assumptions numerically (e.g., mean vs median, skewness), statistically (e.g., KS test) and/or graphically (e.g., box plots, histograms).

The following chapter will now discuss the most common assumptions for statistical test procedures:

- Levene test of variance homogeneity
- Normality test
- Multicollinearity test

That's how it works with DATAtab:

- With DATAtab, checking the assumptions is very quick and easy.
- You will find information about the necessary assumptions directly with the selected test.
- If you want to calculate a regression analysis for example, you open the tab "regression" analysis and choose the variables you want to use for your model.
- Finally, the results are displayed and you find the assumptions check under "Test assumptions" in the results area.

10.1 Levene test of variance homogeneity

Many statistical testing procedures require that there is equal variance in the samples. How can it now be checked whether the variances are homogeneous, i.e. whether there is equality of variance? This is where the Levene test helps. The Levene test checks whether several groups have the same variance in the population.

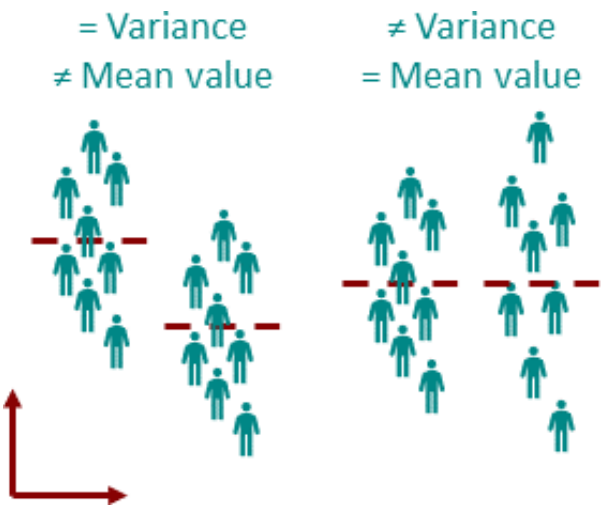


Figure 57: The Levene test of variance homogeneity

Levene's test is therefore used to test the null hypothesis that the samples to be compared come from a population with the same variance. In this case, possible variance differences occur only by chance, since there are small differences in each sampling.

If the p-value for the Levene test is greater than 0.05, then the variances are not significantly different from each other (i.e., the homogeneity assumption of the variance is met). If the p-value for the Levene test is less than 0.05, then there is a significant difference between the variances.

- **H_0:** Groups have equal variances
- **H_1:** Groups have different variances

It is important to note that the mean values of the individual groups have no influence on the result, they may differ. A great advantage of Levene's test is that it is very stable against violations of the normal distribution. Therefore, Levene's test is used in many statistical programs.

The Levene´s test has **two assumptions:**

- independent observations
- test variable has metric scale level

Furthermore, the variance equality can also be checked **graphically.** This is usually done with the help of a **grouped box plot** or a **scatter plot.**

10.2 Levene test example

In this fictitious example, you conducted a survey among students to find out how many cups of coffee they drink per week. Now you want to know whether the variances of the individual subjects are the same and calculate a levene test for this.

Math	History	Psychology
21	18	17
23	22	16
17	19	23
11	26	7
9	13	26
27	24	9
22	23	25
12	17	21
20	21	14
4	15	20

This is how it works with DATAtab:

- To calculate the Levene test, simply copy the above table into the table of the statistics calculator.
- Then click on “t-Test | Chi^2 | ANOVA”.
- Now you simply select the three categories Math, History and Psychology and an ANOVA will be calculated. Here you will now also find a calculated Levene´s test.

As a result, you get two tables and a boxplot. The first table describes the variables descriptively and you can read the standard deviation of each variable.

	N	Mean	Std. Deviation
Math	10	16.6	7.291
History	10	19.8	4.131
Psychology	10	17.8	6.443
Total	30	18.067	6.04

With the help of the **boxplot,** you can visualize the result of the Levene test. The boxplot clearly shows how strongly the variables under investigation scatter.

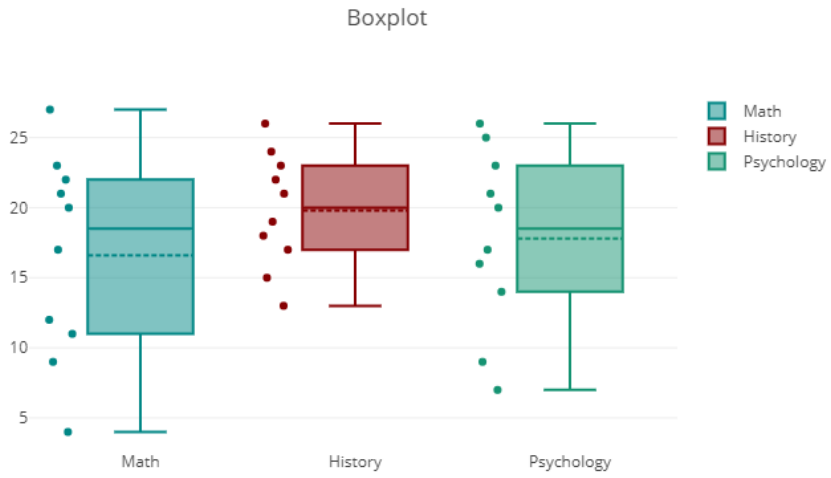


Figure 58: Graphical test of variance homogeneity

After the boxplot you will now get the table with the Levene test statistics. In this table the significance is the most important value, if the significance is above 0.05 there is no difference between the variances of the samples.

Level test of variance equality

F	df1	df2	Significance
2.016	2	27	0.153

Therefore you can easily calculate a Levene test for equality of variances. If the p-value or significance is less than 0.05, you can assume inhomogeneous variance based on the available data.

10.3 Interpreting the Levene Test

The degree of freedom df1 is obtained by calculating the number of groups minus 1, the degree of freedom df2 is obtained by calculating the number of cases minus the number of groups. In this level-test example the significance of 0.153 is greater than the defined significance level of 5%.

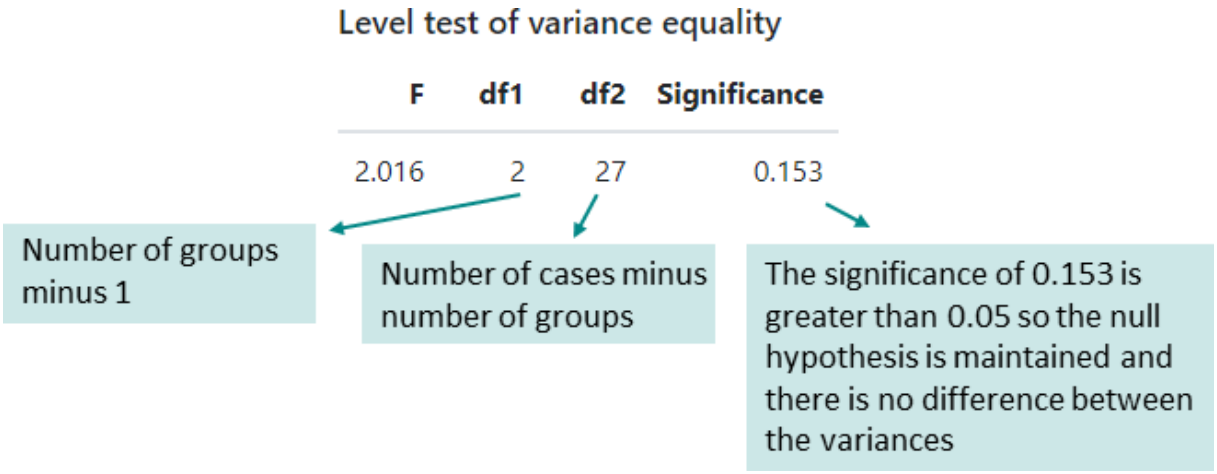


Figure 59: Explanation of the Levene test

Thus the null hypothesis is maintained and there is no difference between the variances of the three groups. Thus, the three samples come from populations with the same variance.

10.4 Normality test

One of the most common assumptions for statistical tests is that the data used are normally distributed. For example, if you want to run a t-test or an ANOVA, you must first test whether the data or variables are normally distributed.

The assumption of normal distribution is also important for linear regression analysis, but in this case it is important that the error made by the model is normally distributed, not the data itself.

Normal distribution can be tested either **analytically (statistical tests) or graphically.** The most common analytical tests to check data for normal distribution are the:

- Kolmogorov-Smirnov Test
- Shapiro-Wilk Test
- Anderson-Darling Test

For graphical verification, either a histogram or, better, the Q-Q plot is used. Q-Q stands for quantile-quantile plot, where the actually observed distribution is compared with the theoretically expected distribution.

10.4.1 Statistical test for normal distribution

To test your data analytically for normal distribution, there are several test procedures, the best known being the Kolmogorov-Smirnov test, the Shapiro-Wilk test, and the Anderson Darling test.

In all of these tests, you are testing the null hypothesis that your data are normally distributed. The null hypothesis is that the frequency distribution of your data is normally distributed. To reject or not reject the null hypothesis, all these tests give you a p-value. What matters is whether this p-value is less than or greater than 0.05.

	Statistics	p-Value
Kolmogorov-Smirnov	0.09	0.892
Shapiro-Wilk	0.98	0.862
Anderson-Darling	0.24	0.772

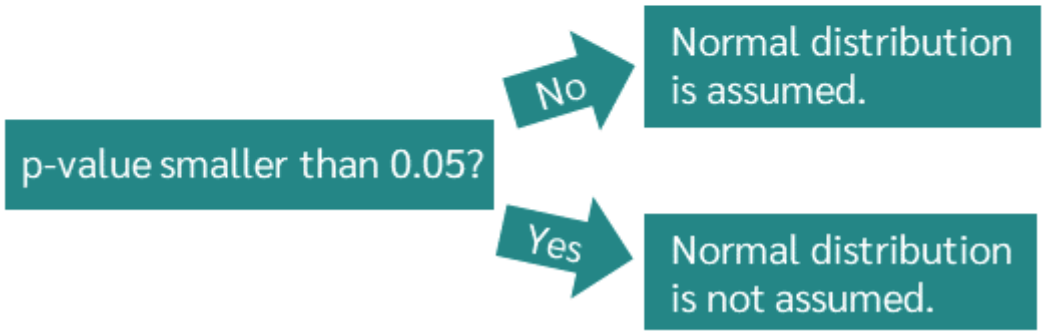


Figure 60: Tests for normal distribution

If the p-value is less than 0.05, this is interpreted as a significant deviation from the normal distribution and it can be assumed that the data are not normally distributed. If the p-value is greater than 0.05 and you want to be statistically clean, you cannot necessarily say that the frequency distribution is normal, you just cannot reject the null hypothesis.

In practice, a normal distribution is assumed for values greater than 0.05, although this is not entirely correct. Nevertheless, the graphical solution should always be considered.

Note: The Kolmogorov-Smirnov test and the Anderson-Darling test can also be used to test distributions other than the normal distribution.

10.4.2 Disadvantage of analytical tests for normal distribution

Unfortunately, the analytical method has a major drawback, which is why more and more attention is being paid to graphical methods.

The problem is that the calculated p-value is affected by the size of the sample. Therefore, if you have a very small sample, your p-value may be much larger than 0.05, but if you have a very large sample from the same population, your p-value may be smaller than 0.05.

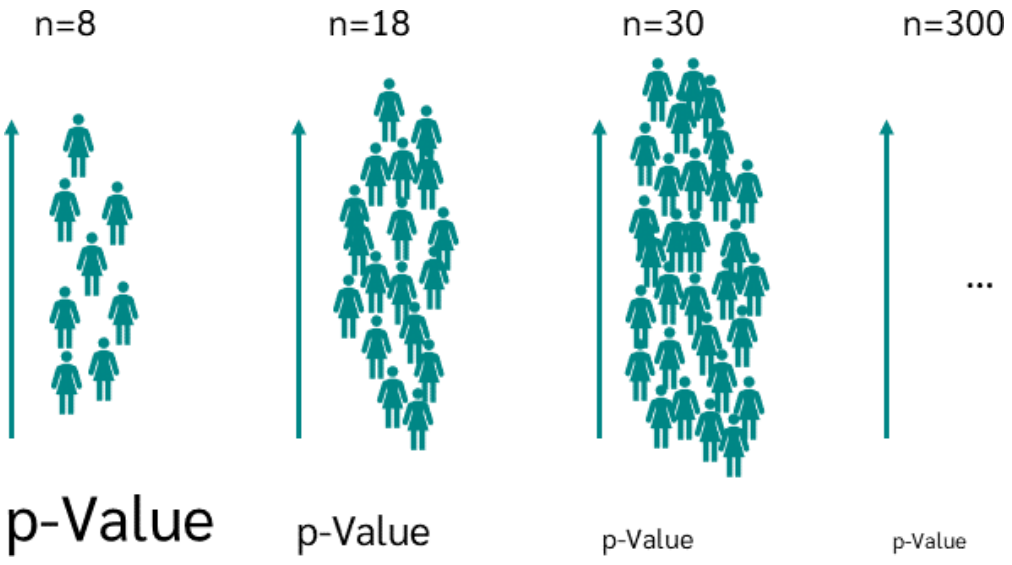


Figure 61: Disadvantage of analytical tests for normal distribution

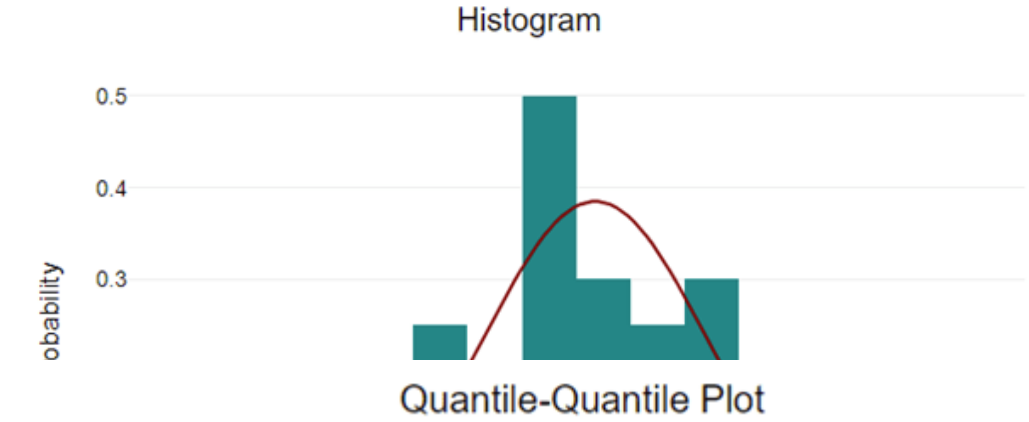
If we assume that the distribution in the population deviates only slightly from the normal distribution, we will get a very large p-value with a very small sample and therefore assume that the data are normally distributed. However, if you take a larger sample, the p-value gets smaller and smaller, even though the samples are from the same population with the same distribution. With a very large sample, you can even get a p-value of less than 0.05, rejecting the null hypothesis of normal distribution.

To avoid this problem, graphical methods are increasingly being used.

10.4.3 Graphical test for normal distribution

If the normal distribution is tested graphically, one looks either at the histogram or even better the QQ plot.

If you want to check the normal distribution using a histogram, plot the normal distribution on the histogram of your data and check that the distribution curve of the data approximately matches the normal distribution curve.



A better way to do this is to use a

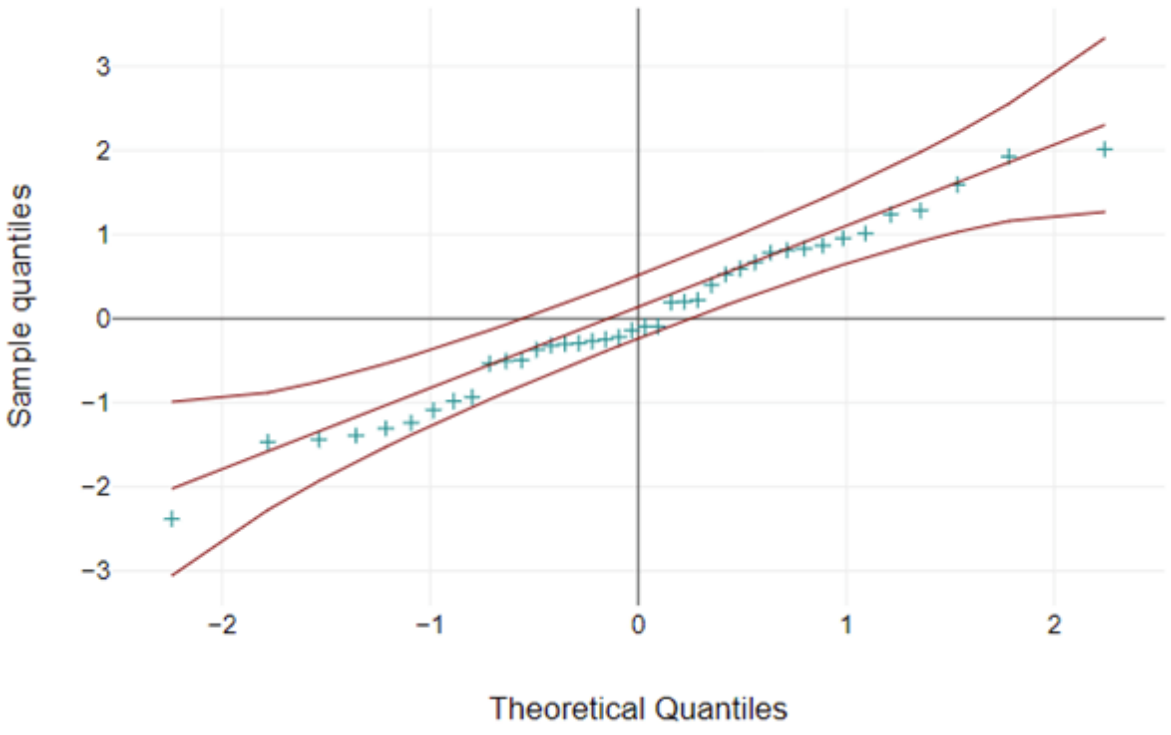


Figure 63: Q-Q plot for testing normal distribution with DATAtab

A better way to do this is to use a quantile-quantile plot, or Q-Q plot for short. This compares the theoretical quantiles that the data should have if they were perfectly normal with the quantiles of the measured values.

If the data were perfectly normally distributed, all points would lie on the line. The more the data deviates from the line, the less the data is normally distributed.

In addition, DATAtab plots the 95% confidence interval. If all or almost all of your data lies within this interval, this is a very strong indication that your data is normally distributed. They would not be normally distributed if, for example, they form an arc and lie far away from the line in some areas.

That's how it works with DATAtab:

When you test your data for normal distribution with DATAtab, you get the following evaluation, first the analytical test procedures clearly arranged in a table, then the graphical test procedures.

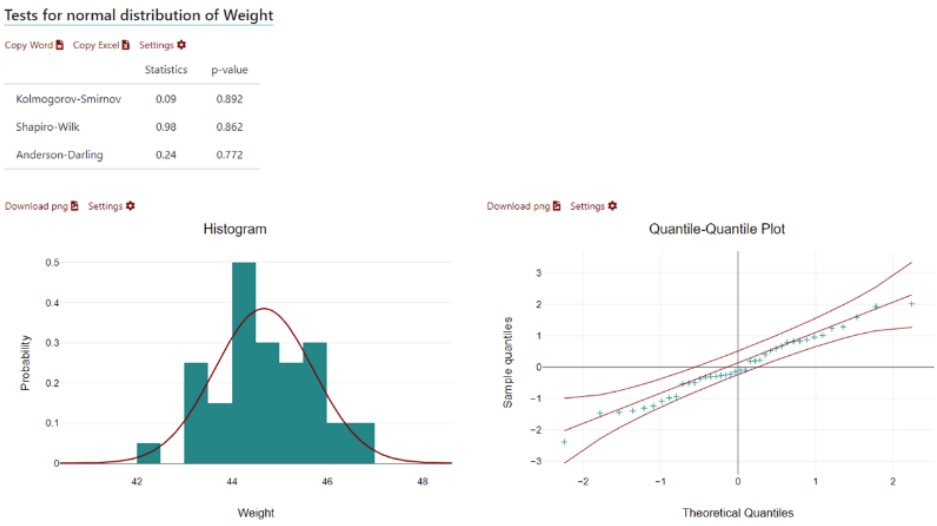


Figure 64: Results of the test for normal distribution with DATAtab

- If you want to test your data for normal distribution, simply copy your data into the table on DATAtab.
- Then click on descriptive statistics and select the variable you want to test for normal distribution.
- Finally, just click on Test Normal Distribution and you will get the results.

Furthermore, if you are calculating a hypothesis test with DATAtab, you can test the assumptions for each hypothesis test, if one of the assumptions is the normal distribution, then you will get the test for normal distribution in the same way.

10.5 Multicollinearity test

In a regression analysis, multicollinearity occurs when two or more predictor variables (independent variables) show a high correlation. This can lead to the regression coefficients being unstable and no longer being interpretable.

Multicollinearity is a problem because it distorts the statistical significance of the independent variable.

A main goal of regression is to determine the relationship of each independent variable and the dependent variable. However, when variables are highly correlated, it may no longer be possible to determine exactly which influence comes from which variable. Thus, the p values of the regression coefficients can no longer be interpreted.

With multicollinearity, the regression coefficients can vary greatly when the data change very slightly or new variables are added.

Multicollinearity only affects the independent variables that are highly correlated. If you are interested in other variables that do not exhibit multicollinearity, then you can interpret them normally.

If you are using the regression model to make a prediction, then multicollinearity does not affect the outcome of the prediction. The multicollinearity only affects the individual coefficients and the p-value.

10.5.1 How to avoid multicollinearity?

To avoid multicollinearity, there must be no linear dependence between the predictors; this is the case, for example, when one variable is the multiple of another variable. In this case, since the variables are perfectly correlated, one variable explains 100% of the other variable and there is no added value in taking both variables in a regression model. If there is no correlation between the independent variables, then there is no multicollinearity.

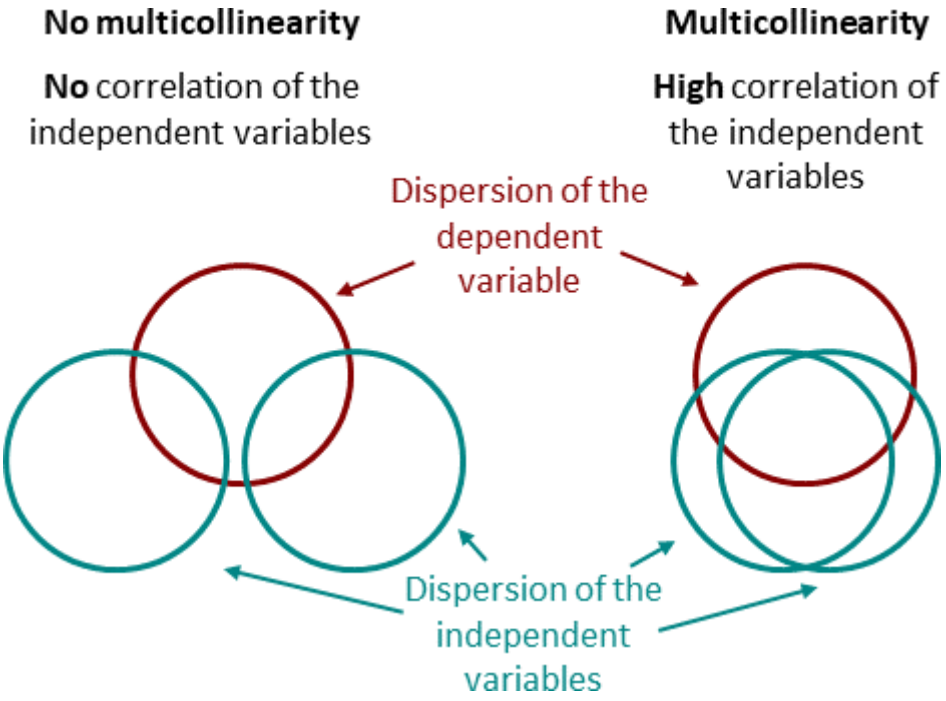

No multicollinearity
No correlation of the independent variables
Multicollinearity
High correlation of the independent variables
Dispersion of the dependent variable
Dispersion of the independent variables

Figure 65: Multicollinearity

In reality, a perfect linear correlation hardly ever occurs, which is why we speak of multicollinearity when individual variables are highly correlated with each other. And in this case the effect of individual variables cannot be clearly separated from each other.

It should be noted that the regression coefficients can no longer be interpreted in a meaningful way, but the prediction with the regression model is possible.

10.5.2 Multicollinearity test

Since there is always some multicollinearity in a given set of data, ratios were introduced to indicate multicollinearity. To test for multicollinearity, a new regression model is created for each independent variable. In these regression models, the original dependent variable is left out and one of the independent variables is made the dependent variable in each case.

Thus, it tests how well one independent variable can be represented by the other independent variables. If the one independent variable can be very well represented by the other independent variables, this is a sign of multicollinearity.

$$\hat{x}_2 = b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a$$

For example, if x_1 can be completely composed of the other variables, then the regression model cannot know what b_1 is or what the other coefficients must be. In mathematics we say that the equation is overdetermined.

10.5.3 Tolerance value

In order to find out whether multicollinearity is present, the tolerance of the individual predictors is considered on the one hand. The tolerance T_i for the *i.* predictor is calculated with

$$T_i = 1 - R_i^2$$

To calculate R_i^2, a new regression model is created, as discussed above. This model contains all predictors, whereby the i. predictor is used as a new criterion (dependent variable). This now makes it possible to determine how well the i. predictor can be represented by the other predictors.

A tolerance value (T) below 0.1 is considered critical and multicollinearity is present. In this case, more than 90% of the variance can be explained by the other predictors.

10.5.4 VIF Multicollinearity

Another measure used to test for multicollinearity is the VIF (Variance Inflation Factor). The VIF statistic is calculated by

$$VIF_i = \frac{1}{1 - R_i^2}$$

The higher the VIF value, the more likely multicollinearity is present. In the VIF test, values above 10 are considered critical. The VIF value therefore increases with increasing multicollinearity.

11. Statistical tests for differences

In the next chapter, statistical tests for testing difference hypotheses are discussed. We will start with parametric tests and explain the one-sample t-test, the independent-samples t-test, and the dependent-samples t-test. In the second part, the Mann-Whitney U test and the Wilcoxon test are discussed, which belong to the nonparametric tests.

11.1 One sample t-test

In this section, the basics of the one-sample t-test will first be discussed and subject-specific questions that can be analyzed with this test procedure will be addressed. The assumptions of the t-test and the differences between one sample and two sample tests are then discussed. Finally, the implementation and interpretation of the results is explained with the help of an example.

11.1.1 Basics of the one sample t-test

The t-test is one of the most common hypothesis tests in statistics. The t-test determines either whether the sample mean and the mean of the population differ or if two sample means differ statistically. The t-test distinguishes between

- One sample t-test
- t-test for independent samples
- t-test for dependent samples

The three variants of the t-test are first illustrated in the following figure and then discussed in more detail.

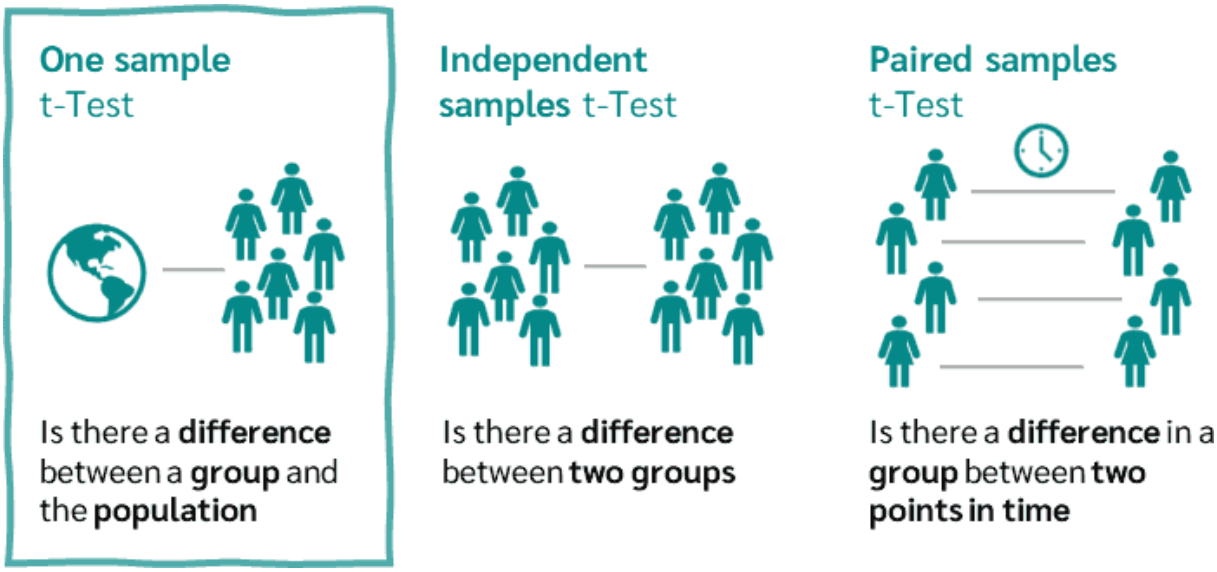

One sample
t-Test
Is there a difference between a group and the population
Independent samples t-Test
Is there a difference between two groups
Paired samples
t-Test
Is there a difference in a group between two points in time

Figure 66: The 3 variants of the t-test

The choice of which t-test to use depends on whether one or two samples are available. If two samples are available, a distinction is made between dependent and independent samples. In this tutorial you will find everything about the one-sample t-test.

Tip: Do you want to calculate the t-value? You can easily calculate it for all three t-tests online in the t-test calculator on DATAtab

The one-sample t-test is used to test whether the population differs from a fixed value. So, the question is: Are there statistically significant differences between a sample mean and the fixed value? The set value may, for example, reflect the remaining population percentage or a set quality target that is to be controlled.

11.1.2 Examples of a t-test for one sample

Let's start with an example from the field of **social sciences** . In this example, you want to find out whether the subjective health perception of managers in Austria differs from that of the population as a whole. To do this, you will ask 50 managers about their subjective perception of health.

Next, we will now discuss a **technical example** , because t-tests are also widely used in technical fields. In this example, you want to find out if the screws your company produces really weigh 10 grams on average. To test this, you

weigh 50 screws and compare the actual weight with the weight they should have (10 grams).

Finally, we discuss an example from the **field of medicine**. A pharmaceutical company promises that its new drug will lower blood pressure by 10 mmHg in one week. You now have to find out whether this is correct. To do this, you compare the observed reduction in blood pressure of 75 subjects with the expected reduction of 10 mmHg.

11.1.3 Assumptions of the one-sample t-test

In a one-sample t-test, the data under consideration must be from a random sample, have metric scale level, and be normally distributed.

One tailed and two tailed t-test

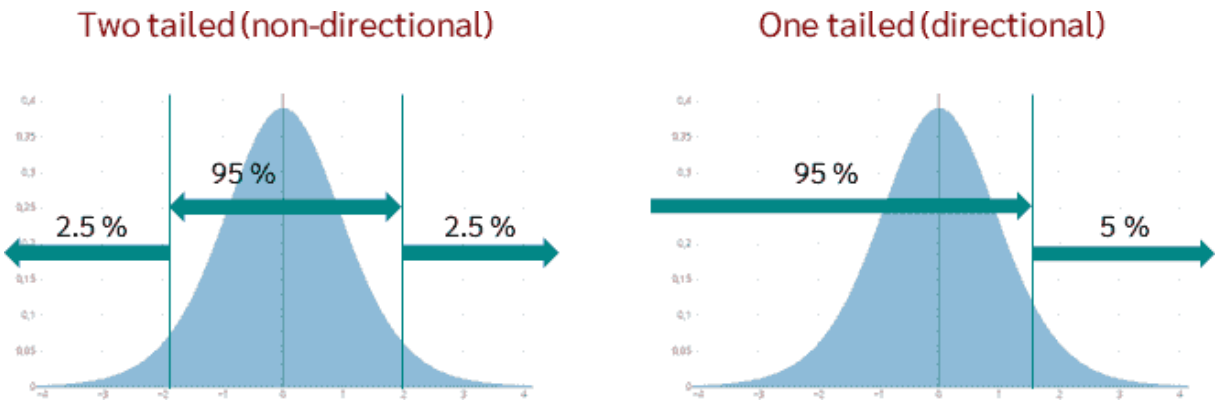


Figure 67: One-sided and two-sided t-test

So if you want to know whether a sample differs from the population, you have to calculate a one sample t-test. But before the t-test can be calculated, a question and the hypotheses must first be defined. This

determines whether a **one-tailed (directional)** or a **two-tailed (non-directional)** t-test must be calculated.

The question helps you to define the object of investigation. In the case of the **one-sample t-test** the question is:

Two-tailed (non-directional)

Is there a statistically significant difference between the mean value of the sample and the population?

One-tailed (directional)

Is the mean value of the sample significantly larger (or smaller) than the mean value of the population?

For the **examples** above, this gives us the following questions:

- Does the health perception of managers in Canada differ from that of the overall population in Canada?
- Does the production plant produce screws with a weight of 10 grams?
- Does the new drug lower blood pressure by 10 mmHg within one week?

11.1.4 Hypotheses for the one-sample t-test

Now, to perform a one-sample t-test in the next step, the following hypotheses are formulated:

two-tailed (non-directional)

- **Null hypothesis H_0**: The mean value of the population is equal to the given value.
- **Alternative hypothesis H_1**: The mean of the population is not equal to the specified value.

one-tailed (directional)

- **Null hypothesis H_0** : The mean value of the population is equal to or greater than (or less than) that of the specified value.
- **Alternative hypothesis H_1** : The mean value of the population is smaller (or larger) than the given value.

11.1.5 Calculation of the one-sample t-test

You can calculate the t-test either with a statistical software like DATAtab or by hand. For the calculation by hand, the **test statistic "t" is** needed first. How the test statistic for the t-test is calculated is shown in the figure below.

Student	Score
1	28
2	29
3	35
4	37
5	32
6	26
7	37
8	39
9	22
10	29
11	36
12	38

Significance level

$$\alpha = 0.05$$

Number of sample values

$$n = 12$$

Mean value

$$\bar{x} = 32.33$$

Standard deviation

$$s = 5.47$$

Standard error of the mean

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{5{,}47}{\sqrt{12}} = 1.58$$

t-value

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{32.33 - 28}{1{,}58} = 2.75$$

Degrees of freedom

$$df = n - 1 = 11$$

Figure 68: t-test statistic

In order to check whether the mean sample value differs significantly from that of the population, the critical t-value must be calculated. First the number of degrees of freedom, abbreviated df, is required, which is calculated by taking the number of samples minus one.

$$\mathrm{df} = N - 1$$

where the standard deviation is the population standard deviation estimated using the sample.

If the number of degrees of freedom is known, the critical t-value can be determined using the table of t-values. For a sample of 12 people, the degree of freedom is 11, and the significance level is assumed to be 5 %. The table

below shows the t values for a one-tailed open distribution. Depending on whether you want to calculate a one-tailed (directional) or two-tailed (non-directional) t test, you must read the t value at either 0.95 or 0.975. For the non-directional hypothesis and an significance level of 5%, the critical t-value is 2.201.

If the calculated t value is below the critical t value, there is no significant difference between the sample and the population; if it is above the critical t value, there is a significant difference.

Table 1: Table of t-values

	Surface one sided										
Degrees of freedom	**0.5**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**	**0.975**	**0.99**	**0.995**	**0.999**	**0.9995**
...	...	...	...	...	...	...	...	...	...	...	...
9	0	0.703	0.883	1.1	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	0	0.7	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	0	0.694	0.87	1.079	1.35	1.771	2.16	2.65	3.012	3.852	4.221
...	...	...	...	...	...	...	...	...	...	...	...

Interpret t-value

The t-value is calculated by dividing the measured difference by the scatter in the sample data. The larger the magnitude of t, the more this argues against the null hypothesis. If the calculated t-value is larger than the critical t-value, the null hypothesis is rejected.

Number of degrees of freedom - df

The number of degrees of freedom indicates how many values are allowed to vary freely. The degrees of freedom are therefore the number of independent individual pieces of information.

11.1.6 One sample t-test with example

As an **example of the one sample t-test,** we examine whether an online statistics tutorial newly introduced at the university has an impact on students' exam scores.

The average score on a college statistics test has been 28 for years. This semester, a new online statistics tutorial was introduced. Now the course director would like to know if student success has changed since the introduction of the statistics tutorial: "Does the online statistics tutorial have a positive effect on exam scores? "

The population considered includes all students who have written the statistics exam since the new statistics tutorial was introduced. The reference value to be compared is 28.

The following null hypothesis (H0) is formulated:

The mean value from the sample and the given value do not differ significantly. The online statistics tutorial has no significant influence on the exam results.

Student	Score
1	28
2	29
3	35
4	37
5	32
6	26
7	37
8	39
9	22
10	29
11	36
12	38

That’s how it works with DATAtab:

Do you want to calculate a t-test independently? Calculate the example in the Statistics Calculator.

- Just copy the upper table including the first row into the t-Test Calculator. Datatab will then provide you with the tables below.
- The following results are obtained with DATAtab: The mean value is 32.33 and the standard deviation 5.46.
- This leads to a standard error of the mean value of 1.57. The t-statistic thus gives 2.75.
- You would now like to know whether your hypothesis (the score is 28) is significant or not.

- To do this, you first specify a significance level in Datatab, usually 5% is used, which is preselected. Now you will get the table below in Datatab.

Statistics

	N	Mean value	Standard deviation	Standard error of the mean value
Score	12	32.33	5.47	1.58

One sample t-test (Test Value = 28)

	t	df	p-value (2-tailed)
Score	2.75	11	0.02

95% confidence interval of the difference

	Mean value difference	Lower	Upper
Score	4.33	0.86	7.81

To interpret whether your hypothesis is significant one of the two values can be used:

- p-value (2-tailed)
- lower and upper confidence interval of the difference

In this example p-value (2-tailed) is equal to 0.02, i.e. 2 %. Put into words this means: The probability that a sample with a mean difference of 4.33 or more will be drawn from the population is 2%. The significance level was set at 5%, which is greater than 2%. For this reason, a significant difference between the sample and the population is assumed.

Whether or not there is a significant difference can also be read from the confidence interval of the difference. If the lower and upper limits go through zero, there is no significant difference. If this is not the case, there is a

significant difference. In this example, the lower value is 0.86 and the upper value is 7.81. Since the lower and upper values do not touch zero, there is a significant difference.

11.1.7 APA format | One-sample t-test

If we were to write the top results for publication in an APA journal, that is, in an APA format, we would write it that way:

A t-test showed a statistically reliable difference between the score of students who attended the online course and the average score of students who did not attend an online course. ($M = 32.33$, $s = 5.47$) and 28, $t(11) = 2.75$, $p < 0.02$, $\alpha = 0.05$.

11.2 T-test for independent samples (unpaired t-test)

The t-test for independent samples (or unpaired t-test) tests whether two independent groups are significantly different.

The t-test for independent samples is used to make a statement about the population based on two independent samples. To do this, the mean value of the two samples is compared. If the difference in the mean values is large enough, it is assumed that the two groups differ.

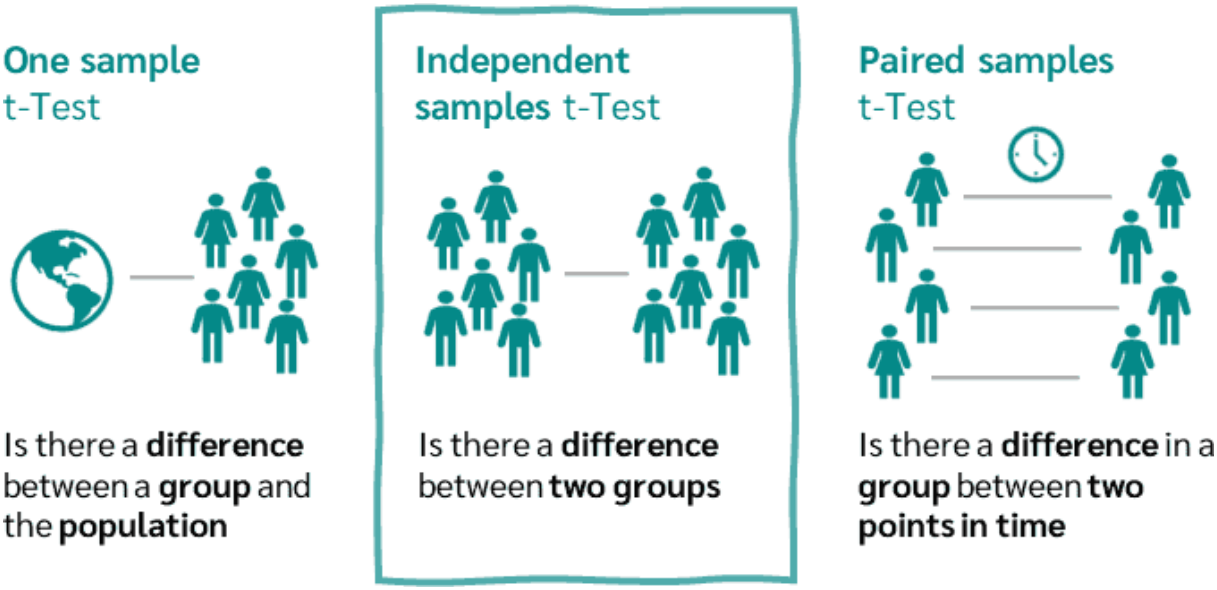

One sample
t-Test
Is there a difference between a group and the population
Independent samples t-Test
Is there a difference between two groups
Paired samples
t-Test
Is there a difference in a group between two points in time

Figure 69: t-test types

11.2.1 Using an independent t-test

Say you want to test if there is a difference between two groups in the population, for example, if there is a difference in salary between men and women. Of course it is not possible to ask all men and women for their salary, so we take a sample. We create a survey and send it randomly to people. In order to be able to make a statement about the population based on this sample, we need the independent t-test.

11.2.2 Purpose of the independent/unpaired t-test

The unpaired t-test puts the mean difference in relation to the standard error of the mean. The standard error of the mean indicates how much the mean value scatter, it indicates how far the sample mean of the data is likely to be from the true population mean. If the fluctuation of the mean value is large, this is an indication that a large difference in the mean values of the two groups is very likely, even by chance.

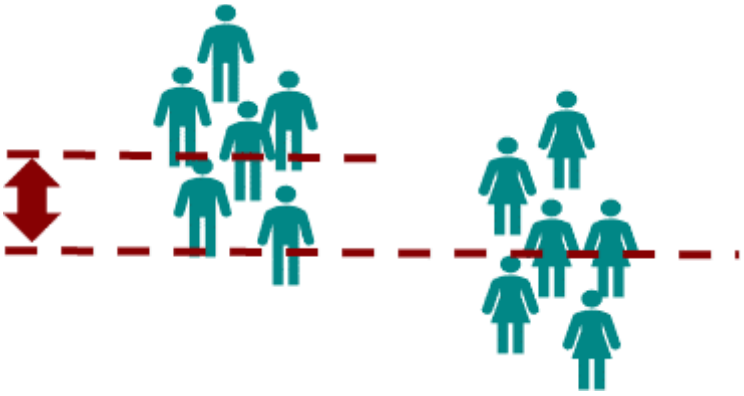


Figure 70: Mean difference

Therefore, the larger the mean difference in the two groups is and the smaller the standard error of the mean, the less likely it is that the given mean difference in the two samples is due to chance.

What are independent samples?

Independent samples exist if no case or person from one group can be assigned to a case or person from the other group. This is the case, for example, when comparing the group of women and the group of men, or the group of psychology students with those of math students.

Paired vs. unpaired t-test

The main difference between the paired and unpaired t-test is the shape of the sample.

- If you have the **same sample** that you survey at two different times, you use a paired t-test.
- If you want to compare **two different groups**, whether they come from one sample or two samples, you use an unpaired t-test.

11.2.3 Examples for the unpaired t-test

There are many applications for the independent or unpaired t-test. For example, it is an important test in the fields of biostatistics or marketing.

Let's start with a **medical example**: For a pharmaceutical company you want to test whether the drug XY helps to lose weight or not. For this purpose, 20 test persons are administered the drug and 20 test persons receive a placebo.

Here is an example from the field of **social sciences**: You want to find out whether there is a difference between people with and without a degree in terms of their health.

Finally, we will discuss a **technical example**: For a screw factory, you want to find out whether two production machines produce screws with the same weight. To test this, you weigh 50 screws from one machine and 50 screws from the other and compare them.

11.2.4 Research question and hypotheses for the unpaired t-test

If you want to know whether two independent groups differ, you must calculate a t-test for independent samples. Before the unpaired t-test can be calculated based on the independent samples, a question must first be formulated, and the hypotheses defined.

With the question you limit your object of investigation. In a t-test for independent samples, the question is generally: *Is there a statistically significant difference between the means of two groups?*

For the examples above, the following questions arise:

- Does drug XY help you lose weight?
- Is there a difference between people with and without a degree in terms of their health?
- Do both production lines produce screws with the same weight?

The next step is to derive the hypotheses to be tested from the research question. Hypotheses are assumptions about reality whose validity is possible but not yet proven.

Two hypotheses are always formulated which assert exactly the opposite. These are the null hypothesis and the alternative hypothesis.

Null hypothesis H_0	**Alternative hypothesis H_1**
There is no mean difference between the two groups in the population → Two population means are equal. → The two groups are from the same population → H_0: $\mu_1 = \mu_2$ **Example:** There is **no difference** between the salary of men and women	There is a mean difference between the two groups in the population → The two population means are not equal → The two groups are not from the same population → H_1: $\mu_1 \neq \mu_2$ **Example:** There is **a difference** between the salary of men and women

11.2.5 Assumptions unpaired/independent t-test

To calculate an independent t-test, there must be an independent variable (e.g., gender) that has two characteristics or groups (e.g., male and female). These two groups are to be compared in the analysis. The question is, is there a difference between the two groups with regard to the dependent variable (e.g. income).

The assumptions are now the following:

1. Two groups or samples must be independent

As the name of this t-test suggests, the samples must be independent. This means that a value in one sample must not have any influence on a value in the other sample.

- Correct: Measure the weight of people who have dieted and people who have not dieted.
- Wrong: Measure the weight of a person before and after a particular diet.

2. Dependent variable must be metric

In the t-test for independent samples, the mean value of the sample must be calculated. This only makes sense if the variable is metrically scaled.

- Correct: The weight of a person (in kg).
- Wrong: A person's level of education (University, High School, ...).

3. Variables must be normally distributed

The t-test for independent samples provides the most accurate results if the data of the groups are each normally distributed. However, there are exceptions to this in special cases.

- Correct: The weight, age or height of a person.
- Wrong: The number of points after rolling a die (uniform distribution, since the probability of each point is 1/6).

4. The variance within the groups should be similar

Since the variance is needed for the test statistic t, it must be the same within the groups.

- Correct: Weight, age or height of a person.
- Wrong: Stock market prices in "normal" times and in a recession.

Assumptions not met?

If the assumptions for the independent t-test are not met, the calculated p-value may be incorrect. However, if the two samples are of equal size, the t-test is quite robust to a slight skewness of the data. The t-test is not robust if the variances differ significantly.

If the variables are not normally distributed, the Mann-Whitney U test can be used. The Mann-Whitney U Test is the non-parametric counterpart of the independent t-test.

If the variables are **not normally distributed**, the Mann-Whitney U- test can be used. The Mann-Whitney U test is the **nonparametric counterpart of** the independent t test.

11.2.6 Calculate t-test for independent samples

Depending on whether the variance between the two groups is assumed to be equal or unequal, a different formula for the test statistic "t" results. The test of whether the variances are equal or not is done with **Levene's test**. The null hypothesis in Levene's test is that the two variances are not different. Thus, if the Levene test results in a p-value of less than 5%, it is assumed that there is a difference in the variances of the two groups.

Formula for equal (homogeneous) variance

If Levene's test results in a p-value greater than 5%, both groups are assumed to have the same variance, and the test statistic for the unpaired t-test is given by

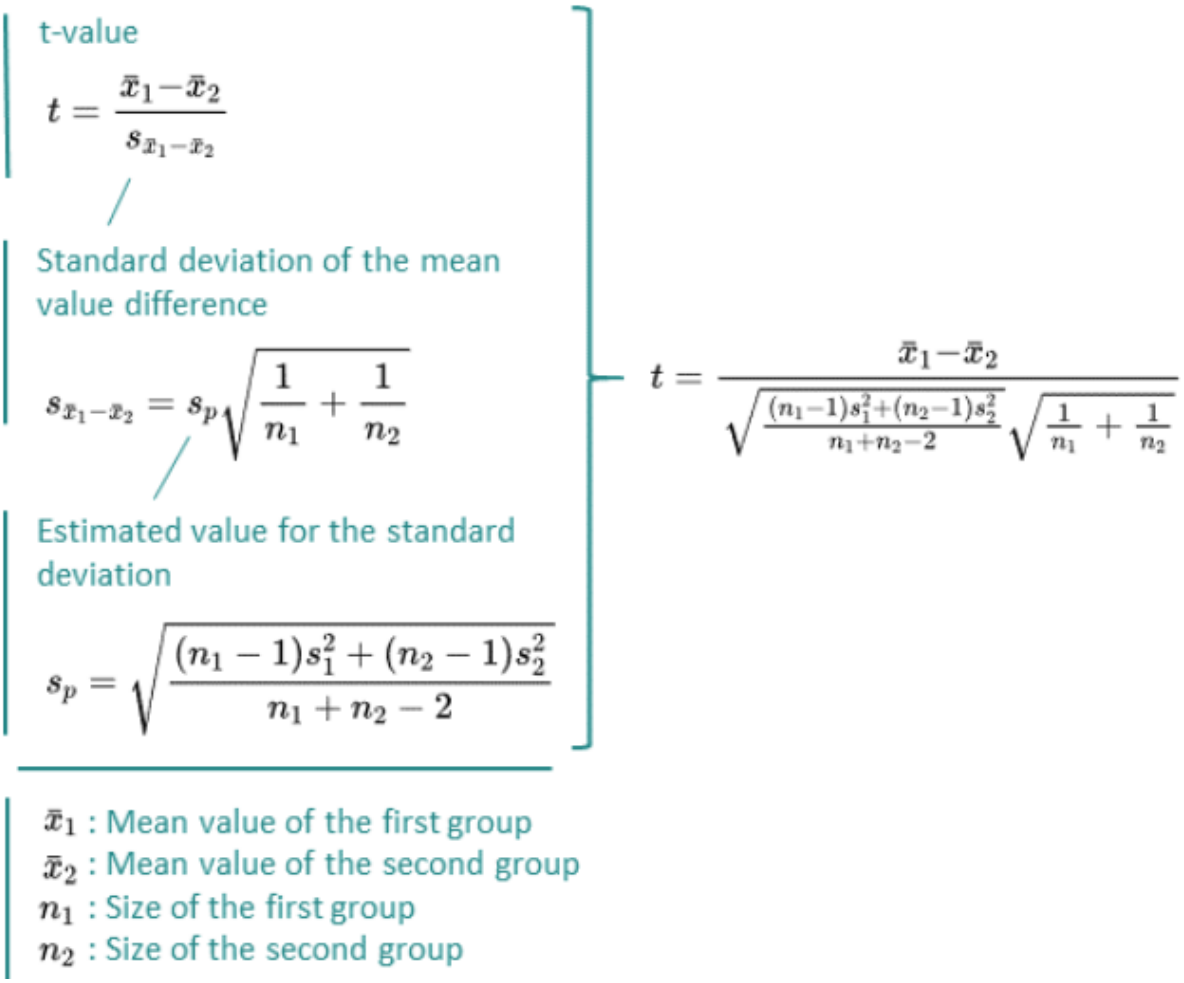

t-value
$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$
Standard deviation of the mean value difference
$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Estimated value for the standard deviation
$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
$\bar{x}_1$: Mean value of the first group
$\bar{x}_2$: Mean value of the second group
n_1 : Size of the first group
n_2 : Size of the second group

Figure 71: Calculation of the t-test for independent samples.

The critical p-value can then be determined from the table with the t-distribution. The degree of freedom results with

$$df = n_1 + n_2 - 2$$

,

where n1 and n2 again indicate the number of cases in the two samples.

Formula for unequal (heterogeneous) variance

The test statistic t for a t-test for independent samples with unequal variance is calculated via

t-value

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$: Mean value of the first group
$\bar{x}_2$: Mean value of the second group
n_1 : Size of the first group
n_2 : Size of the second group
s_1 : Standard deviation of the first group
s_2 : Standard deviation of the second group

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1 - 1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2 - 1}(\frac{s_2^2}{n_2})^2}$$

Figure 72: Calculation of the t-value

The p-value then follows from the table with the t-distribution, where the degrees of freedom are distributed over

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1 - 1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2 - 1}(\frac{s_2^2}{n_2})^2}$$

result.

11.2.7 Confidence interval for the true mean difference

The calculated mean difference in the independent t-test has been calculated using the sample. Now it is of interest in which range the true mean difference lies. To determine within which limits the true difference is likely to lie, the confidence interval is calculated.

The 95% confidence interval for the true mean difference can be calculated by the following formula:

$$KI = \bar{x}_1 - \bar{x}_2 \pm t^* \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where t* is the t value obtained at 97.5% and degrees of freedom df.

11.2.8 One-sided and two-sided unpaired t-test

As explained in the chapter on hypothesis testing, there are one-sided and two-sided hypotheses (also called directed and undirected hypotheses). To be fair, there is also a one-sided and two-sided t-test for independent samples. By default, the two-sided unpaired t-test is calculated, which is also output in DATAtab.

To obtain the one-sided t-test for independent samples, the p-value must be divided by two. Now it depends on whether the data "tend in the direction" of the hypothesis or not. If the hypothesis states that the mean value of one group is larger or smaller than the mean value of the other group, this must also be seen in the result. If this is not the case, one minus the halved p-value must be calculated.

11.2.9 Effectsize unpaied t-test

The effect size for an unpaired t-test is usually calculated using the **hedges g,** or simply called ***d***.

In the unpaired t-test calculator on DATAtab you can easily display the effect strength.

What do you need the effect size for? The calculated p-value depends very much on the sample size. For example, if there is a difference in the population, the larger the sample size, the more clearly this difference is "indicated" in the p-value. Thus, if the sample size is chosen very high, even very small differences, which may no longer be relevant, can be "detected" in the population. To standardize this, the effect size is used in addition to the p-value.

11.2.10 Example t-test for independent samples

In this example, a lecturer wants to know whether the statistics exam results in the summer semester differ from those in the winter semester. To do this, she creates an overview with the points achieved per exam.

Accordingly, the research **question** is "Is there a significant difference between exam scores in the summer and winter semesters? "

The **null hypothesis H0** is formulated as follows: There is no difference between the two samples. There is no difference between the statistics exam results in the summer semester and the winter semester.

Finally, the **alternative hypothesis H1** is: The two samples differ from each other. There is a difference between the statistics exam results in the summer semester and the winter semester.

Semester	Points
Summer	52
Summer	61
Summer	40
Summer	46
Summer	50
Summer	56
Summer	44
Summer	47
Summer	70
Summer	40
Summer	65

Semester	Points
Summer	38
Summer	68
Winter	53
Winter	71
Winter	38
Winter	34
Winter	68
Winter	68
Winter	46
Winter	41
Winter	38
Winter	23
Winter	28

That’s how it works with DATAtab:

After copying the above sample data into the hypothesis testing calculator, you can output the t test for independent samples. The results for the t test example look like this:

Hypotheses

Copy Word Copy Excel

Null hypothesis	Alternative hypothesis
There is no difference between the Summer term and Winter term groups with respect to the dependent variable Points	There is a difference between the Summer term and Winter term groups with respect to the dependent variable Points

Descriptive statistics

Copy Word Copy Excel

		N	Mean	Std. Deviation	Std. Error Mean
Points	Summer term	13	52.08	11.03	3.06
	Winter term	11	46.18	16.71	5.04

Levene test of variance equality

Copy Word Copy Excel

F	df1	df2	p
2.44	1	22	.133

t-Test for independent samples

Copy Word Copy Excel

		t	df	p (2-tailed)
Points	Equal variances	1.04	22	.312
	Unequal variances	1	16.82	.331

95% Confidence Interval of the Difference

Copy Word Copy Excel

		Mean Difference	Standard Error of Difference	Lower limit	Upper limit
Points	Equal variances	5.9	5.69	-5.92	17.71
	Unequal variances	5.9	5.89	-6.55	18.34

11.2.11 Interpretation t-test for independent samples

To make a statement about whether your hypothesis is significant or not, one of the following two values is used:

- p-value (two-sided)
- lower and upper confidence interval of the difference

In this example t-test, the p-value (two-sided) is 0.312, or 31%. This means that the probability that you draw a sample where both groups differ more than the groups in the example is 31%. Since the significance level has been set at 5%, this makes it much lower than 31%. For this reason, no significant difference is assumed between the two samples and therefore they come from the same population.

The second way to determine whether or not there is a significant difference is to use the confidence interval of the difference. If the lower and upper limits pass through zero, there is no significant difference. If it does not, there is a significant difference. In this example of an unpaired t-test, the lower value is -6.328 and the upper value is 18.118. Since the lower and upper values graze zero, there is no significant difference.

It is common practice to first display the two independent samples in a diagram before calculating a t-test for independent samples. A boxplot is suitable for this purpose, which visualizes the position measures and scatter measures of the two independent samples very well.

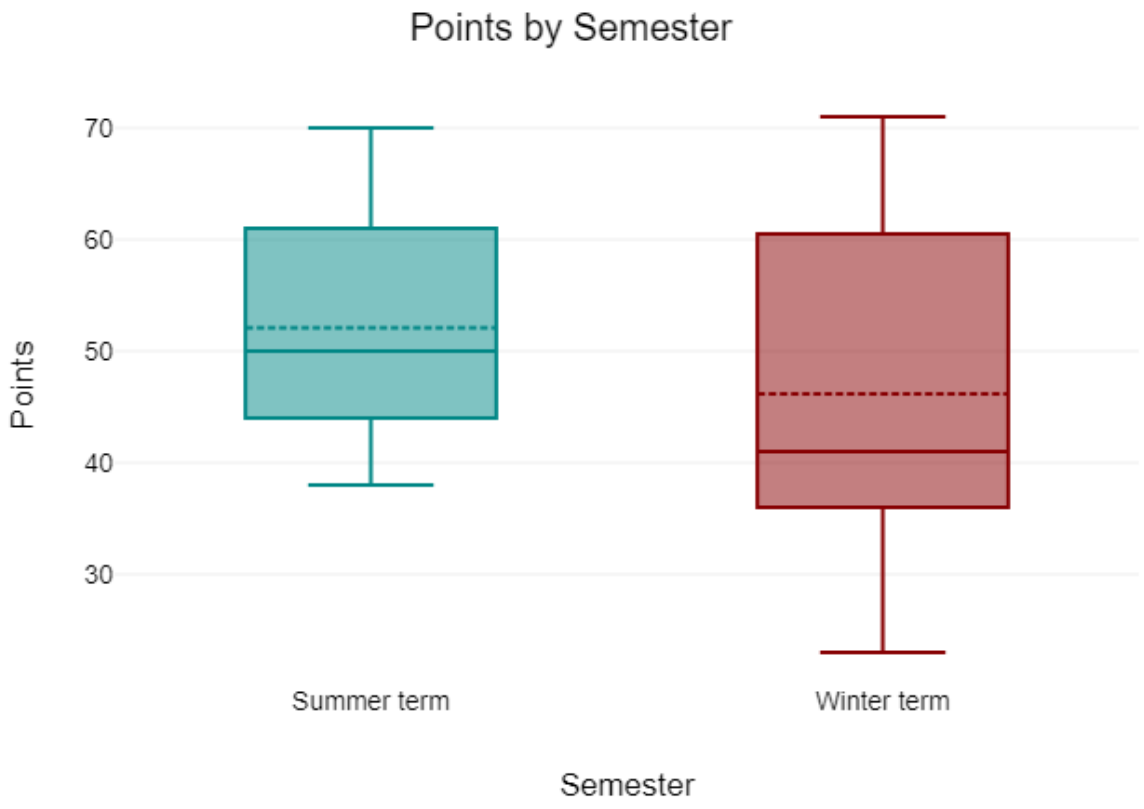


Figure 73: Boxplot showing the t-test results.

11.2.12 Report a t-test for independent samples

Reporting a t-test for independent samples in APA (American Psychological Association) style involves presenting key details about your statistical test in a clear, concise manner. Here's a general guideline on how to report the results of an independent samples t-test according to APA style:

- Test Statistic:

Clearly state that you are using an independent samples t-test. Report the degrees of freedom in parentheses after the "t" statistic, then provide the value of t.

- Significance Level:

This is typically reported as "p" followed by the exact value or a comparison

- Effect Size:

It's good practice to include an effect size (like Cohen's d) alongside the t-test result. This provides an indication of the magnitude of the difference between groups.

Means and Standard Deviations:

Report the means and standard deviations for each group. This gives a context to the t-test result.

Sample Size:

You can also mention the number of participants in each group, especially if this wasn't previously stated.

Here's a template:

An independent samples t-test was conducted to compare [variable] in [group 1] and [group 2]. There was a significant difference in the scores for [group 1] (M = [mean], SD = [standard deviation]) and [group 2] (M = [mean], SD = [standard deviation]); t([degrees of freedom]) = [t value], p = [exact p value] (two-tailed). The magnitude of the differences in the means (mean difference = [mean difference], 95% CI: [lower limit, upper limit]) was [small, medium, large], with a Cohen's d of [d value].

For example, consider you conducted an independent samples t-test comparing test scores between males and females. Assume you found the following results:

- Males: M = 50, SD = 10, n = 30
- Females: M = 55, SD = 9, n = 30
- t(58) = -2.5, p = .015, Cohen's d = 0.5

The results would be reported as:

An independent samples t-test was conducted to compare test scores in males and females. There was a significant difference in the scores for males (M = 50, SD = 10) and females (M = 55, SD = 9); t(58) = -2.5, p = .015 (two-tailed). The magnitude of the differences in the means (mean difference = -5,

95% CI: [provide the confidence interval limits here]) was medium, with a Cohen's d of 0.5.

11.3 Paired-samples t-test

The paired-samples t-test (t-test for dependent samples) is a statistical test used to determine if there is a difference between two dependent groups.

The t-test for dependent samples, or paired t-test, tests whether the mean values of two dependent groups differ significantly from each other.

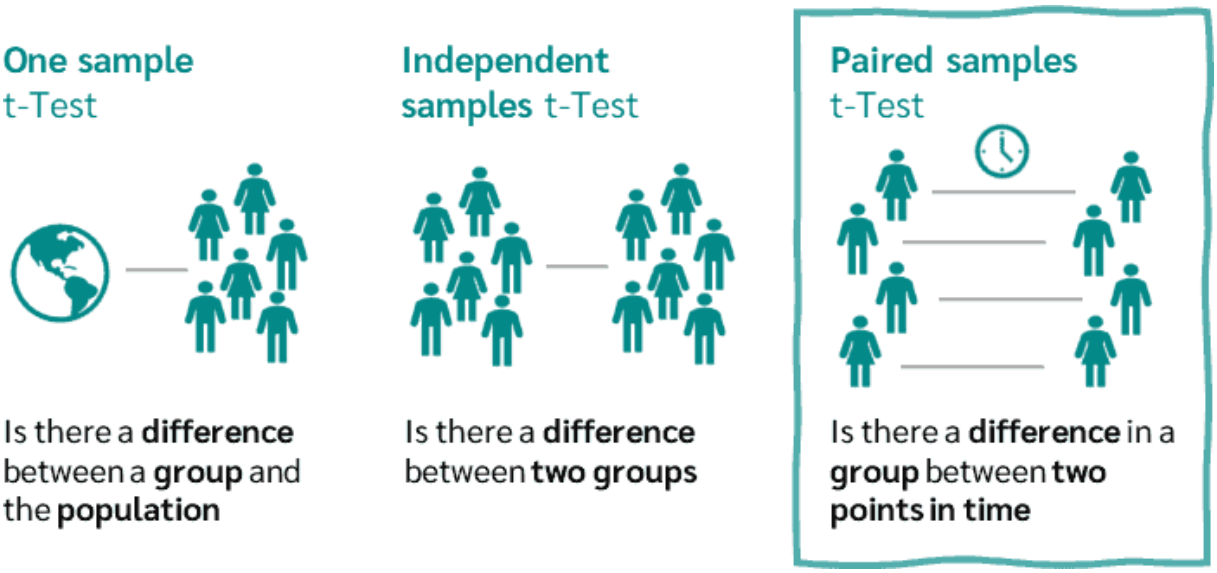

One sample
t-Test
Is there a difference between a group and the population
Independent samples t-Test
Is there a difference between two groups
Paired samples
t-Test
Is there a difference in a group between two points in time

Figure 74: Forms of the t-test

11.3.1 Why do you need the paired t-test?

You need the paired t-test whenever you survey the same group or sample at two points in time. For example, you might be interested in whether a rehabilitation program has a positive effect on physical fitness. Since you can't

ask all the people who go to rehab, you use a random sample. You can then use the paired t-test to infer the population from the sample.

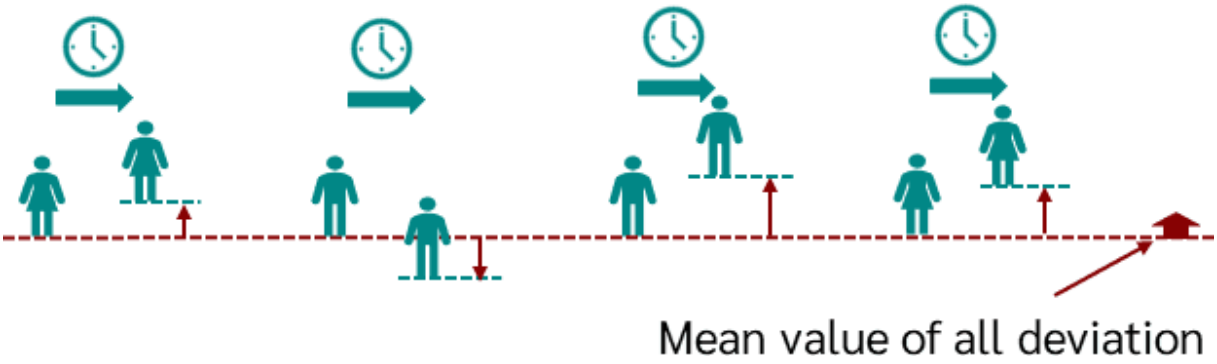


Figure 75: t-test for dependent and paired samples, respectively.

In **dependent samples**, the measured values are available in pairs. These pairs result from repeated measurements, parallelization or matching. This can be the case, for example, in longitudinal studies with several measurement points (time series analyses) or in intervention studies with experimental designs (before-after measurement).

An example of **dependent sampling** would be to measure the weight of a group of people at two points in time. A person can then be unambiguously assigned a weight at the first and second measurement time and the difference between the measured values can be calculated in each case.

If more than two measurement times are available, ANOVA with repeated measures is used.

11.3.2 What is the advantage of a dependent t-test over an independent t-test?

The question of whether to use a dependent t-test or an independent t-test is, of course, already determined as part of the study design, and it is not possible to arbitrarily use either one test or the other. Therefore, the question is rather which type of study makes more sense:

- Conducting a study with one group of participants who are measured twice.
- To conduct a study with two separate groups of participants, each measured once.

The major advantage of a repeated-measures design that then uses the paired t-test is that individual differences between participants can be eliminated. This means that the probability of detecting a (statistically significant) difference, if one exists, is higher with the paired t-test than with the independent t-test.

11.3.3 Examples of the t-test for paired samples

The t-test for dependent samples has numerous applications, here are three examples.

Medical example:

In a pharmaceutical company, you want to test whether a new drug increases memory performance. To do this, you determine the memory performance of 40 test subjects before and after they have taken the drug.

Technical example:

A screw factory complains about very high downtimes in its five production lines. You now want to find out whether a newly introduced lubricant has an influence on the downtimes. To do this, you compare the downtimes of the five plants before and after the introduction of the new lubricant.

Social science example:

You want to find out whether there has been a change in the German population's health awareness between 2010 and 2015. To do this, you could use data from the Socio-Economic Panel (SOEP), for example. The SOEP is a representative repeat survey of private households in Germany. The same people are always surveyed at regular intervals on the same topics. To answer your question, compare the health awareness of respondents in 2010 and 2015.

11.3.4 Research question and hypotheses of the paired t-test

In order to calculate a t-test for dependent samples, a research question and the hypotheses must first be defined.

In a t-test for dependent samples, the question is generally: Is there a statistically significant difference between the sample mean of two dependent groups?

The questions for the above examples arise as follows:

- Does the new drug help to increase memory performance?
- Does the newly introduced lubricant have an impact on downtimes?
- Has the health awareness of the German population changed between 2010 and 2015?

Now the hypothesis can be derived from the question. In the hypothesis, a provisional, i.e., not certain, assumption is made that is to be tested. In a paired t-test, the hypotheses are:

- **Null hypothesis H_0**: The means of the two dependent groups are equal.
- **Alternative hypothesis H_1**: The means of the two dependent groups differ.

11.3.5 Assumptions paired t-test

Of course, before calculating the dependent t-test, the assumptions still must be checked. These are now described in the following section. If the assumptions 2 and 3 are not fulfilled, the Wilcoxon test must be used. The Wilcoxon test is the non-parametric counterpart of the paired t-test.

1. Two dependent groups or samples are available

As the name *t-test for dependent samples* implies, the groups must be dependent, i.e., a value of one group must belong to a value of the other group.

- From one and the same person the weight is measured before and after a diet. (correct)
- The weight of people who have dieted and people who have not dieted is measured. (wrong)

2. The variables are metric scaled

In the t-test for dependent samples, the difference between the two dependent values is formed and then the mean is calculated. This only makes sense if the values are metric.

- The salary of a person (in €). (correct)
- The educational qualification of a person (secondary school, high school, etc.). (wrong)

3. The differences of the paired values is normally distributed

The difference between the paired values required for the paired t-test must be normally distributed.

- The difference in the weight of a person at two points in time. (correct)

- The difference of the number of points after throwing two dice. (wrong)

11.3.6 Calculating a paired t-test

In the paired t-test, the difference is calculated from each paired case. The mean value is then calculated from these differences. Depending on how large the mean value is and how large the standard error of the mean value is, a statement is then made as to how likely it is that this result has arisen by chance.

For the calculation of the t-test for dependent samples, the difference of each pair from the two groups is first formed. From the resulting differences, the mean value x̄diff is then calculated.

The calculation of the test statistics t is now equal to the t test for one sample. If there is no difference between the two groups, the mean value of the difference x̄diff is zero. So the question is, is there a difference between x̄diff and zero.

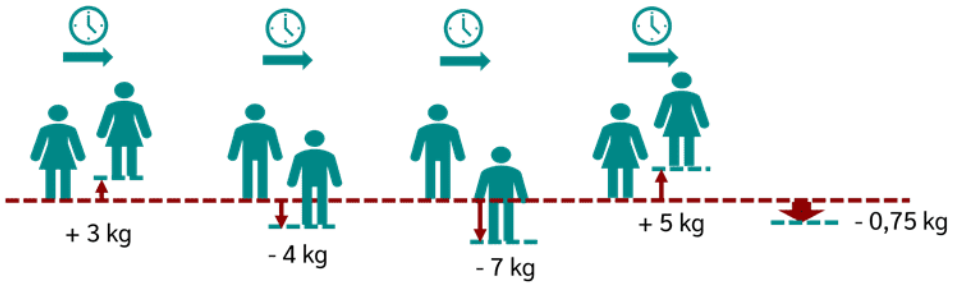
The test statistic t for the t-test for dependent samples is then calculated as

$$t = \frac{\bar{x}_{diff} - 0}{s_{\bar{x}}}$$

where $s_{\bar{x}}$ is the standard error of the mean value

$$s_{\bar{x}} = \frac{s_{diff}}{\sqrt{N}}$$

- x_{diff} = Difference between the groups
- $\bar{x}_{diff}$ = Mean value of the difference between the two groups
- N = Sample size
- s_{diff} = Standard deviation
- $s_{\bar{x}}$ = Estimated standard error of mean



Number of cases

$$n = 4$$

Degrees of freedom

$$df = n - 1 = 3$$

Mean

$$\bar{x} = \frac{+3 - 4 - 7 + 5}{4} = -0.75$$

Standard deviation

$$s = \sqrt{\frac{(3 + 0.75)^2 + (-4 + 0.75)^2 + (-7 + 0.75)^2 + (5 + 0.75)^2}{4 - 1}} = 5.68$$

Standard error of the mean

$$s_e = \frac{s}{\sqrt{n}} = \frac{5.68}{2} = 2.84$$

t-Value

$$t = \frac{\bar{x} - 0}{s_e} = \frac{\bar{x}}{s_e} = \frac{-0.75}{2.84} = -0.26$$

11.3.7 Example t-test for dependent samples with DATAtab

In the dependent/paired samples t-test example, we examine whether summer vacation has an impact on students' physical fitness.

Thus, the research question is: "Do summer vacations have an influence on the physical fitness of statistics students? "To test this, ten statistics students will be given a fitness test once before and once after the vacations (two measurement points).

The null hypothesis is formulated as follows: The mean difference of the pairs of measurements (before and after the vacations) is equal to zero. The semester break has no influence on the physical fitness of the students.

Now, since two test results always come from one student, there is a dependency between the two samples. Therefore, the t-test for dependent samples is calculated.

The table of test results looks as follows:

Statistics Student	Score before vacations	Score after vacations
1	60	61
2	70	71
3	40	38
4	41	39
5	40	38
6	40	33
7	45	55
8	48	56
9	30	38
10	50	68

That's how it works with DATAtab:

After copying the upper table into the t-test calculator you can calculate the paired t-test. The results look like this:

Descriptive statistics

Copy Word Copy Excel

	N	Mean	Std. Deviation	Std. Error Mean
Score before vacations	10	46.4	11.45	3.62
Score after vacations	10	49.7	14.1	4.46

Box plot

Orientation
- (•) vertical
- ( ) horizontal

Show Points
- ( ) yes
- (•) no

Standard Deviation
- ( ) yes
- (•) no

Size of the graphic
- ( ) small
- (•) medium
- ( ) large
- ( ) extra large

Download png Download svg Settings

Correlation

Copy Word Copy Excel

	N	Correlation	p
Score before vacations - Score after vacations	10	0.85	.002

t-Test for paired samples

Copy Word Copy Excel

	t	df	p (2-tailed)
Score before vacations - Score after vacations	-1.39	9	.197

95% Confidence Interval of the Difference

Copy Word Copy Excel

	Mean	Std. Deviation	Std. Error Mean	Lower limit	Upper limit
Score before vacations - Score after vacations	-3.3	7.5	2.37	-8.66	2.06

11.3.8 Interpretation of a t-test for dependent samples

If the calculated p-value is smaller than the specified significance level (usually 5%), the null hypothesis is rejected, otherwise it is retained. For the upper example, you can report the results as follows:

The score of the pre-vacation variable had lower values (*M* = 46.4, *SD* = 11.452) than the score of the post-vacation variable (*M* = 49.7, *SD* = 14.095). A dependent samples t-test showed that this difference was not statistically significant: *t*(9) = -1.392, *p* = 0.197, 95% confidence interval [-8.664, 2.064].

This results in a p-value of 0.197 which is above the defined significance level of 0.05. The t-test result is therefore not significant, and the null hypothesis is not rejected.

11.3.9 Effect size dependent t-test

The indication of the effect size is very important for empirical studies. To make a statement about the effect size in a t-test for dependent samples, the following equation can be used:

$$Cohen's\ d = \frac{\bar{x}_{diff}}{s}$$.

In general, it can be said about the effect size:

- Effect size *d*: 0.2 small effect
- Effect size *d*: 0.5 medium effect
- Effect size *d*: 0.8 large effect

11.4 Mann-Whitney U test

The Mann-Whitney U-Test can be used to test whether there is a difference between two samples (groups), and the data need not be normally distributed.

To determine if there is a difference between two samples, the rank sums of the two samples are used rather than the means as in the t-test for independent samples.

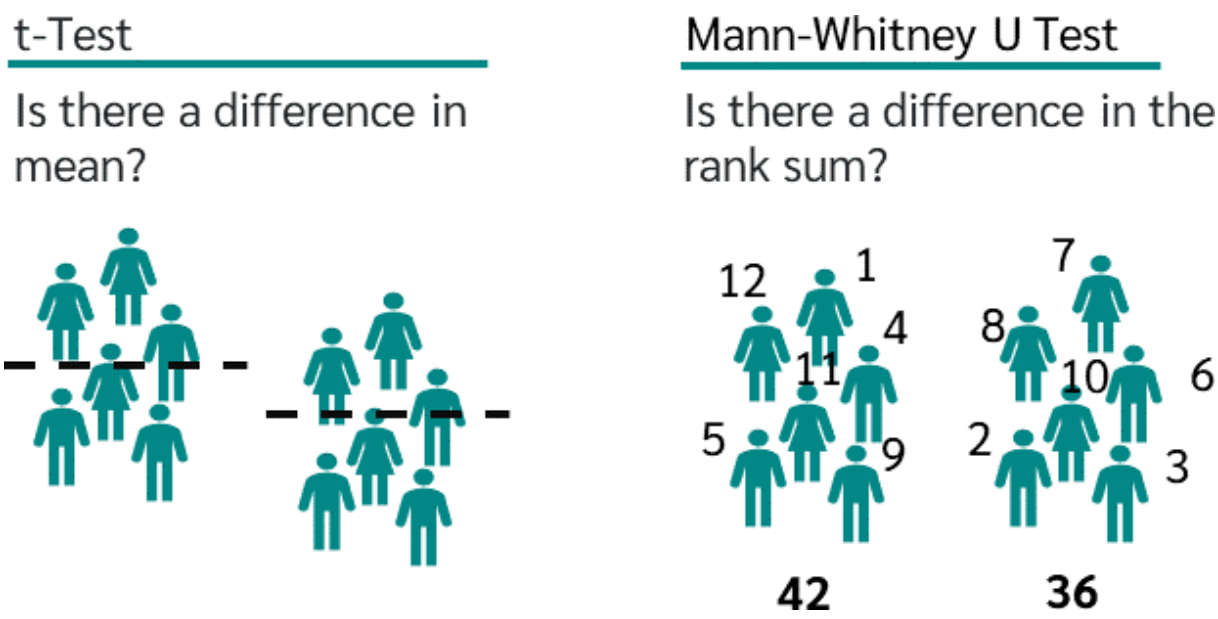


Figure 76: t-test and Mann-Whitney U-test

The Mann-Whitney U test is thus the **non-parametric** counterpart to the t test for independent samples. It is subject to less stringent requirements than the t-test. Therefore, the Mann-Whitney U test is always used when the requirement of normal distribution for the t test is not met.

11.4.1 Assumptions Mann-Whitney U test

To compute a Mann-Whitney U test, only two independent samples with at least ordinal scaled characteristics need to be available. The variables do not have to satisfy any distribution curve.

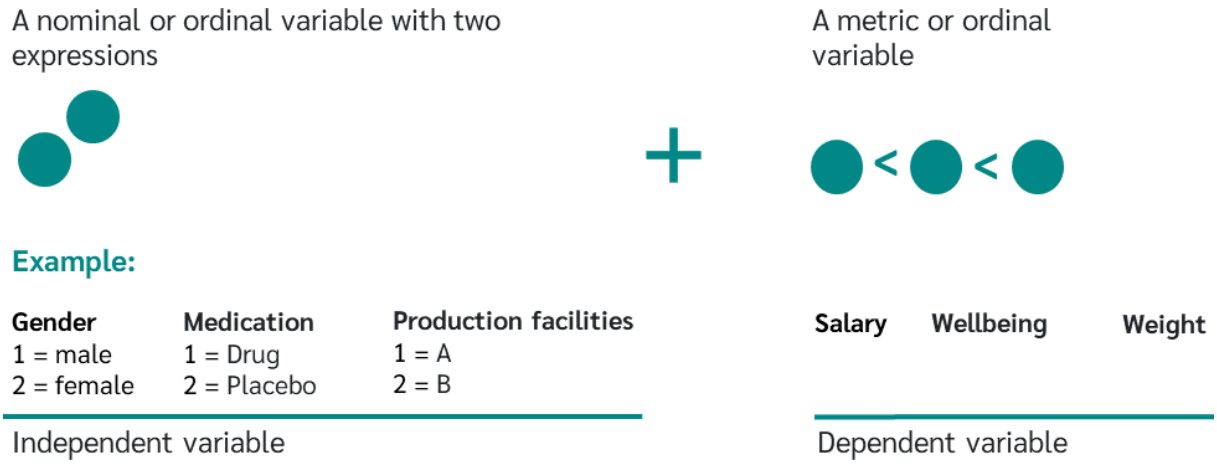


Figure 77: Assumptions of the U-test

If the data are available in pairs, the Wilcoxon test must be used instead of the Mann-Whitney U test.

11.4.2 Hypotheses Mann-Whitney U test

The hypotheses of the Mann-Whitney U test are very similar to the hypotheses of the independent t-test. The difference, however, is that in the case of the Mann-Whitney U test, the test is based on a difference in the central tendency, whereas in the case of the t test, the test is based on a difference in the mean values. Thus, the Mann-Whitney U test results in:

- Null hypothesis: There is no difference (in terms of central tendency) between the two groups in the population.
- Alternative hypothesis: There is a difference (with respect to the central tendency) between the two groups in the population.

11.4.3 Calculate Mann-Whitney U test

To calculate the Mann-Whitney U test for two independent samples, the rankings of the individual values must first be determined.

Gender	Reaction time	Rang
female	34	2
female	36	4
female	41	7
female	43	9
female	44	10
female	37	5
male	45	11
male	33	1
male	35	3
male	39	6
male	42	8

Calculation of the rank sums

$$T_1 = 2 + 4 + 7 + 9 + 10 + 5 = 37$$

$$T_2 = 11 + 1 + 3 + 6 + 8 = 29$$

Figure 78: Calculate rank sums

These rankings are then added up for the two groups. In the example above, the rank sum T_1 of the women is 37 and the rank sum T_2 of the men is 29. The average value of the rank places is thus $\bar{R}_1$ = 6.17 for women and $\bar{R}_2$ = 5.80 for men. The difference between $\bar{R}_1$ and $\bar{R}_2$ now shows whether there are possible differences between the reaction times.

In the next step, the U-values are calculated from the rank sums T_1 and T_2

Female

Number of cases $n_1 = 6$ Rank sum $T_1 = 37$

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - T_1$$
$$= 6 \cdot 5 + \frac{6 \cdot (6 + 1)}{2} - 37$$
$$= 14$$

Male

Number of cases $n_2 = 5$ Rank sum $T_2 = 29$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - T_2$$
$$= 6 \cdot 5 + \frac{5 \cdot (5 + 1)}{2} - 29$$
$$= 16$$

U-Wert

$$U = min(U_1, U_2) = min(14, 16) = 14$$

Expected value of U

$$\mu_U = \frac{n_1 \cdot n_2}{2} = \frac{6 \cdot 5}{2} = 15$$

Standard error of U

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{6 \cdot 5 \cdot (6 + 5 + 1)}{12}} = 5.4772$$

z-value

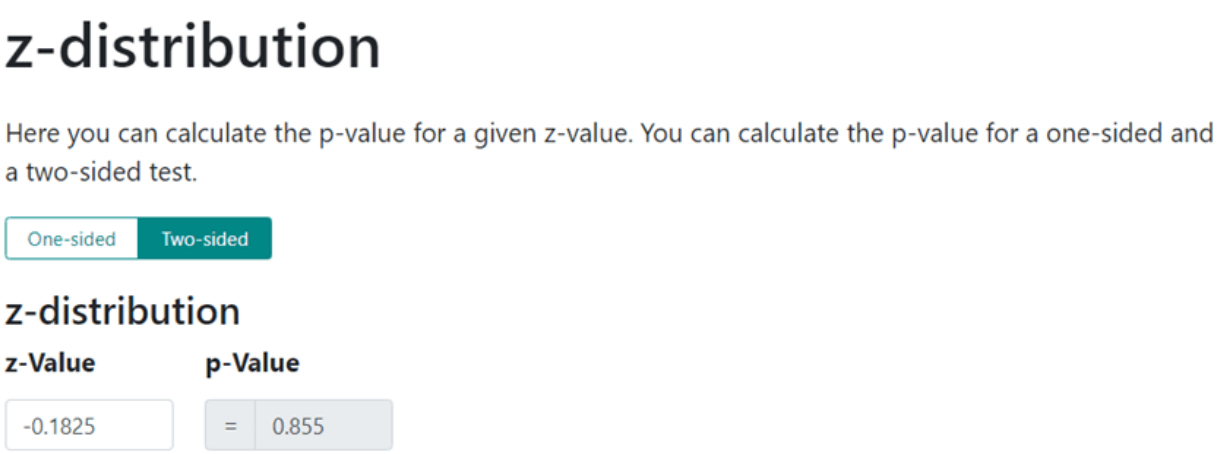
$$z = \frac{U - \mu_U}{\sigma_U} = \frac{14 - 15}{5{,}4772} = -0.1825$$

where n_1, n_2 are the number of elements in the first and second group, respectively. If both groups come from the same population, i.e., the groups do not differ, then the expected value of U is obtained for both U values. After the mean and dispersion have been estimated, **z** can now be calculated. For the Mann-Whitney U value, the smaller value of U_1 and U_2 is used.

Depending on how large the sample is, the **p-value** for the Mann-Whitney U test is calculated in different ways. For up to 25 cases, the exact values are used, which can be read from a table. For larger samples, the normal distribution can be used as an approximation.

Note: In this example, we would actually use the **exact value**, but we will still use the normal distribution. For this, you simply insert the z-value into the **"z-value to p-value calculator"** of DATAtab.

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{14 - 15}{5.4772} = -0.1825$$



If the calculated z-value is larger than the critical z-value, the two groups differ.

11.4.4 Calculate Mann-Whitney U test with tied ranks

If several people share a rank, connected ranks are present. In this case, there is a change in the calculation of the rank sums and the standard deviation of the U-value. We will now go through both using an example.

In the example it can be seen that the...

- ...reaction times 34 occur twice and share the ranks 2 and 3
- ...reaction times 39 occur three times and share the ranks 6, 7 and 8.

Gender	Response time	Ranks
female	33	1
female	**34**	**2 and 3** → $\frac{2+3}{2} = 2.5$
male	**34**	
male	36	4
male	37	5
female	**39**	
female	**39**	**6, 7 and 8** → $\frac{6+7+8}{3} = 7$
male	**39**	
male	43	9
male	44	10
female	45	11

Gender	Ranks
female	1
female	**2.5**
male	**2.5**
male	4
male	5
female	**7**
female	**7**
male	**7**
male	9
male	10
female	11

To account for these connected ranks, the mean values of the joined ranks are calculated in each case. In the first case, this results in a "new" rank of 2.5 and in the second case in a "new" rank of 7. Now the rank sums T can be calculated.

Gender	Ranks
female	1
female	2.5
male	2.5
male	4
male	5
female	7
female	7
male	7
male	9
male	10
female	11

Calculation of the rank sums

$$T_{female} = 1 + 2.5 + 7 + 7 + 11 = 28.5$$

$$T_{male} = 2.5 + 4 + 5 + 7 + 9 + 10 = 37.5$$

Number of tied ranks

Number of people sharing rank i

$$\sum_{i=1}^{k} \frac{t_i^3 - t_i}{12} = \frac{2^3 - 2}{12} + \frac{3^3 - 3}{12} = 2.5$$

Since the rank ties are clearly visible in the upper table, a term is calculated here that is needed for the later calculation of the u-value in the presence of rank ties.

Now all values are available to calculate the z-value considering connected ranks.

Female

Number of cases	Rank sum
$n_1 = 5$	$T_1 = 28.5$

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - T_1$$
$$= 5 \cdot 6 + \frac{5 \cdot (5 + 1)}{2} - 28.5$$
$$= 16.5$$

Male

Number of cases	Rank sum
$n_2 = 6$	$T_2 = 37.5$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - T_2$$
$$= 5 \cdot 6 + \frac{6 \cdot (6 + 1)}{2} - 37.5$$
$$= 13.5$$

Number of all cases

$$n = n_1 + n_2 = 11$$

U-value

$$U = \min(U_1, U_2) \quad = \min(16{,}5\ , 13{,}5) \ = 13.5$$

Expected value of U

$$\mu_u = \frac{n_1 \cdot n_2}{2} \quad = \frac{6 \cdot 5}{2} = 15$$

Standard error of U

$$\sigma_{U_{corr}} = \sqrt{\frac{n_1 \cdot n_2}{n \cdot (n-1)}} \cdot \sqrt{\frac{n^3 - n}{12} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{12}}$$

$$\sigma_{U_{corr}} = \sqrt{\frac{5 \cdot 6}{11 \cdot (11-1)}} \cdot \sqrt{\frac{11^3 - 11}{12} - 2{,}5} \ = 5.41$$

z-value

$$z = \frac{U - \mu_U}{\sigma_{u_{corr}}} \quad = \frac{13.5 - 15}{5.41} \quad = -0.28$$

Again, noting that you actually need about 20 cases to assume normal distribution of u values.

11.4.5 Mann-Whitney U test Example with DATAtab

A Mann-Whitney U Test can be easily calculated with DATAtab.

Simply copy the table below or your own data into the statistics calculator and click on Hypothesis tests.

Then click on the two variables and select Non-Parametric Test.

Gender	Reaction time
female	34
female	36
female	41
female	43
female	44
female	37
male	45
male	33
male	35
male	39
male	42

DATAtab then gives you the following table for the Mann-Whitney-U test:

Hypotheses

Copy Settings

Null hypothesis	Alternative hypothesis
There is no difference between the female and male groups with respect to the dependent variable Response time	There is a difference between the female and male groups with respect to the dependent variable Response time

Descriptive statistics

Copy Settings

		N	Mean	Median	Standard deviation
Response time	female	6	39.167	39	4.07
	male	5	38.8	39	4.919

Box plot

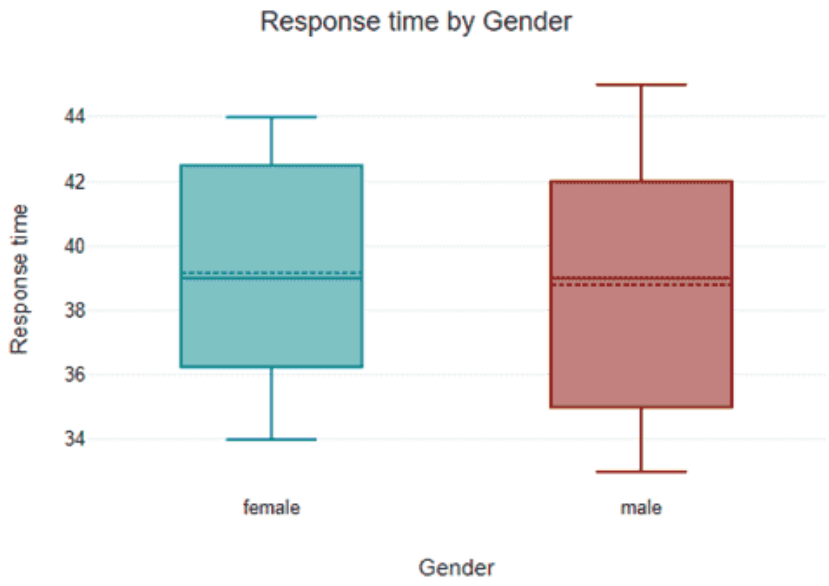
Orientation
- (•) vertical
- ( ) horizontal

Show Points
- ( ) yes
- (•) no

Standard Deviation
- ( ) yes
- (•) no

Size of the graphic
- ( ) small
- (•) medium
- ( ) large
- ( ) extra large

Download png Download svg Settings



Ranks

Copy Settings

	N	Mean Rank	Sum of Ranks
female	6	6.167	37
male	5	5.8	29
Total	11		

Statistics for Mann-Whitney U-Test

Copy Settings

	Values
Mann-Whitney U	14
Z	-0.183
Asymptotic Significance (2-tailed)	.855
Exact Significance (2-tailed)	.931

The **Mann-Whitney U test** works with ranks; therefore, the **mean ranks** and the **rank sum are** displayed first in the result. The reaction time of the women has a slightly lower value than that of the men.

Ranks

Copy Settings

	N	Mean Rank	Sum of Ranks
female	6	6.167	37
male	5	5.8	29
Total	11		

Statistics for Mann-Whitney U-Test

Copy Settings

	Values
Mann-Whitney U	14
Z	-0.183
Asymptotic Significance (2-tailed)	.855
Exact Significance (2-tailed)	.931

DATAtab gives you the **asymptotic significance** and the **exact significance**. Which significance is used depends on the **sample size.** As a rule:

- $n_1 + n_2 < 30$→ exact significance
- $n_1 + n_2 > 30$→ asymptotic significance

Therefore, exact significance is used for this example. The **significance (two-sided)** is 0.931, which is above the significance level of 0.05. Therefore, no difference between the reaction time of men and women can be detected with these data.

11.4.6 Interpret Mann-Whitney U test

The reaction time female group had the same high values (Mdn= 39) as the reaction time male group (Mdn= 39). A Mann-Whitney U-Test showed that this difference was not statistically significant, U=14, p=.931, r=0.06.

11.4.7 Mann-Whitney U test and effect size

n order to make a statement about the Effect Size in the Mann-Whitney-U-Test, you need the Standardised test statistic z and the number of pairs n, with this you can then calculate the Effect Size with the equation below

$$r = \frac{|z|}{\sqrt{n}}$$

In this case, an effect size r of 0.06. In general, one can say about the effect strength:

- Effect size *r* less than 0.3 small→ effect
- Effect size *r* between 0.3 and 0.5→ medium effect
- Effect size *r* greater than 0.5→ large effect

11.5 Wilcoxon test

The Wilcoxon test (Wilcoxon signed-rank test) determines whether two dependent groups differ significantly from each other. To do this, the Wilcoxon test uses the ranks of the groups instead of the mean values.

The Wilcoxon test is a non-parametric test and therefore has fewer assumptions than its parametric counterpart, the paired samples t-test. Thus, when the assumptions for the dependent samples t-test are not met, the Wilcoxon test is used instead.

Medical example:

Load Data

Comparing Pain Levels Before and After Treatment: A study measures patients' pain (1–10 scale) before and after medication. Since pain scores may not be normally distributed, the Wilcoxon Signed-Rank Test compares pre- and post-treatment levels.

Technical example:

Load Data

Battery Life Before and After a Software Update: An engineer tests battery life on the same set of devices before and after a software update. Since battery performance may not be normally distributed, the Wilcoxon Signed-Rank Test determines if the update significantly affects battery life.

11.5.1 Assumptions of the Wilcoxon test

The Wilcoxon Signed-Rank Test is a great non-parametric alternative to the paired t-test, especially when the normality assumption is violated. However, this assumption must be met to perform a Wilcoxon test:

- **Repeated measurement:** The test is used for paired or dependent samples, meaning the same subjects or units are measured before and after an intervention or under two different conditions. So a characteristic of a person, e.g. weight, was measured at two points in time
- **Metric or Ordinal Data:** The data should be at least ordinal or metric (e.g., measurements like pain levels, reaction times, or weights). It cannot be used for nominal (categorical) data.
- **Symmetric Distribution of Differences:** The differences between paired observations should be symmetrically distributed around the median. Unlike the paired t-test, the Wilcoxon test does not require normality but works best when the distribution is roughly symmetric.
- **No Significant Outliers in the Differences:** Extreme outliers can affect the ranking process, reducing the reliability of the test.
- **Random Sampling:** The sample should be randomly selected from the population to ensure unbiased results.

If the data are not available in pairs, the Mann-Whitney U test is used instead of the Wilcoxon test.

11.5.2 Hypotheses in the Wilcoxon test

The hypotheses of the Wilcoxon test are very similar to the hypotheses of the dependent t-test. However, in the case of the Wilcoxon test, the test is whether there is a difference in the central tendency; in the case of the t-test, the test is whether there is a difference in the mean. Thus, the Wilcoxon test test results in:

Null hypothesis: There is no difference (in terms of central tendency) between the two groups in the population.

Alternative hypothesis: There is a difference (with respect to the central tendency) between the two groups in the population.

11.5.3 Wilcoxon test and test power

Now of course the question may come, why don't I just always use the Wilcoxon test instead of the t-test for dependent samples? Then I don't need to test for normal distribution! Parametric tests like the t-test are usually more powerful!

With a parametric test, a smaller difference or a smaller sample is usually enough to reject the null hypothesis. Both are, of course, very convenient. Therefore, if possible, always use parametric tests!

11.5.4 Calculate Wilcoxon test

To perform the Wilcoxon test for two dependent samples, first, calculate the differences between the paired values. Then, take the absolute values of these differences and rank them accordingly. It is crucial to retain the original signs of the differences throughout the process. (An example with tied ranks follows.)

Reaction time morning	Reaction time evening	diff (morning - evening)	Ranking from |diff|
34	45	-11	6 (-)
36	33	3	2
41	35	6	5
39	43	-4	3 (-)
44	42	2	1
37	42	-5	4 (-)
			$T^+ = 8$ & $T^- = 13$

In each case the sum of positive and negative rankings

Figure 79: Calculation of the Wilcoxon test

In the final step, the rank sums are computed separately for the positive and negative differences.

Sum of positive ranks

$$T^+ = 2 + 5 + 1 = 8$$

$$T^- = 6 + 3 + 4 = 13$$

Sum of negative ranks

The test statistic W is then calculated using the sum of positive ranks. Note that there are different ways to calculate W; sometimes, the maximum or minimum value of T+ and T− is used for W.

$$W = T^+ = 8$$

Test statistic W

In this example, the test statistics *W* results in 13. If there is no difference between the two dependent samples, the expected value can be calculated using the following formula:

Number of non-zero differences

$$u_W = \frac{n\,(n+1)}{4} = \frac{6\,(6+1)}{4} = 10.5$$

Expected value of W

Next, the test statistic W is compared to the expected value by calculating the standardized test statistic z. For this we need the Standard deviation.

Standard deviation

$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$= \sqrt{\frac{6(6+1)(2\cdot 6+1)}{24}}$$

$$= 4.78$$

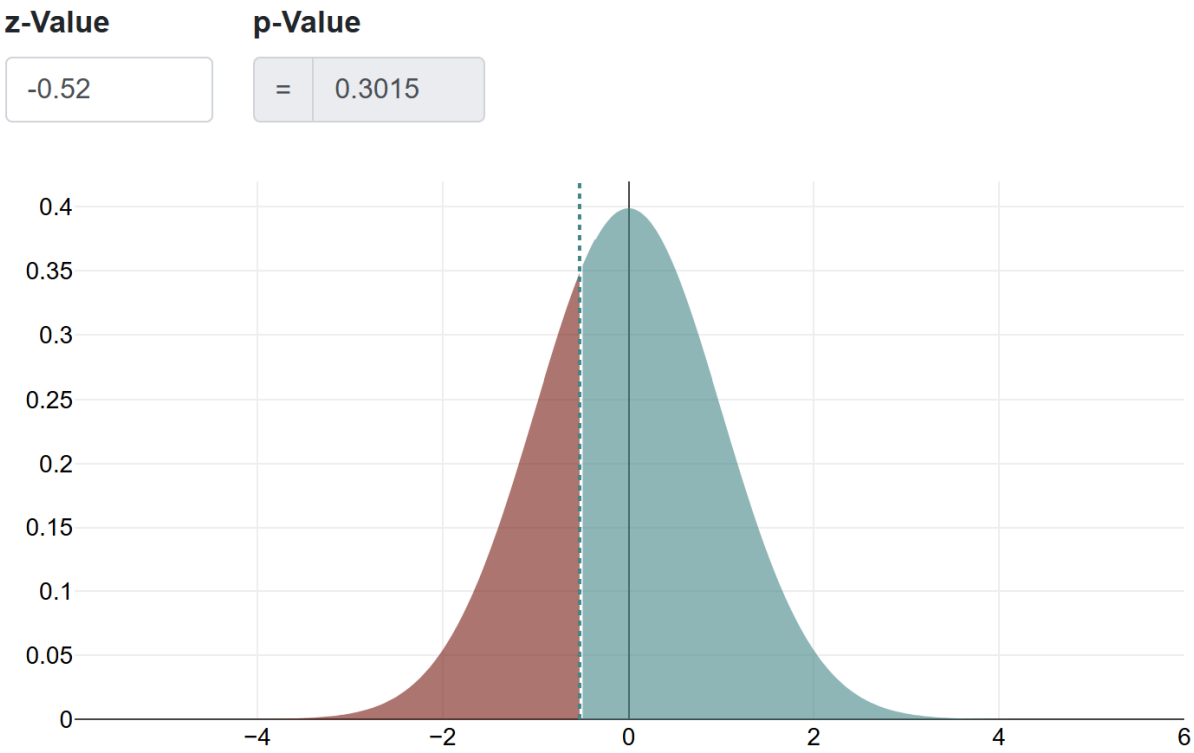
Now we have everything needed to calculate the z value.

$$z = \frac{W - \mu_W}{\sigma_W}$$

$$= \frac{8 - 10.5}{4.78}$$

$$= -0.52$$

We can now determine whether there is a significant difference between the two groups by calculating the corresponding p-value for the z statistic.



Note: This approach is typically valid when the sample size is greater than 25, ensuring the distribution approximates normality.

Since we are testing a two-sided hypothesis, we multiply the p-value of 0.3 by 2, resulting in a final p-value of 0.6.

Continuity Correction

Many statistical software programs, such as DATAtab, apply a so-called continuity correction in the normal approximation for the p-value. As a result, the p-value may vary slightly.

Wilcoxon-Test

Copy AI Interpretation

	W	z	p	r
Reaction time morning - Reaction time evening	8	-0.52	.675	0.21

And here is the entire calculation workflow presented in a single figure:

Total ranks

$$T^+ = 2 + 5 + 1 \quad = 8$$

$$T^- = 6 + 3 + 4 \quad = 13$$

Test statistics W

$$W = T^+ = 8$$

Expected value of W

$$u_W = \frac{n\,(n+1)}{4} = \frac{6\,(6+1)}{4} = 10.5$$

Standard deviation

$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}} = 4.78$$

z-Value

$$\mathrm{z} = \frac{W - \mu_W}{\sigma_W} = \frac{8 - 10.5}{4.78} = -0.52$$

11.5.5 Calculate Wilcoxon signed-rank test with tied ranks

If several people share a rank, connected ranks are present. In this case, there is a change in the calculation of the rank sums and the standard deviation of the W-value. We will now go through both using an example.

In the example it can be seen that there are...

- ...three people who have a difference in amount of two, these people share the ranks 2, 3 and 4.
- ...two people who have a difference in amount of 4, these people share the ranks 6 and 7.
-

morning	evening	diff	Rank |diff|			Ranks
43	44	-1	1			1 (-)
36	38	-2				
43	41	2	2, 3 and 4	→	$\frac{2+3+4}{3} = \underline{3}$	**3** (-)
41	39	2				**3**
37	34	3	5			3
37	41	-4				5
43	39	4	6 and 7	→	$\frac{6+7}{2} = \underline{6.5}$	**6.5** (-)
40	34	6	8			**6.5**
						8

To account for these connected ranks, the mean values of the joined ranks are calculated in each case. In the first case, this results in a "new" rank of 3 and in the second case in a "new" rank of 6.5. Now we can calculate the rank sums of the positive and negative ranks.

Ranks	Rank sums
1 (-)	$T^- = 1 + 3 + 6.5 = 10.5$
3 (-)	
3	$T^+ = 3 + 3 + 5 + 6.5 + 8 = 25.5$
3	
5	
6.5 (-)	
6.5	
8	

Number of tied ranks

Number of people sharing rank i

$$\sum_{i=1}^{k} \frac{t_i^3 - t_i}{12} = \frac{3^3 - 3}{2} + \frac{2^3 - 2}{2} = 15$$

Since the rank ties are clearly visible in the upper table, a term is calculated here that is needed for the later calculation of the W-value in the presence of rank ties.

Now all values are available to calculate the z-value considering connected ranks.

Total ranks

$$T^- = 1 + 3 + 6.5 = 10.5$$
$$T^+ = 3 + 3 + 5 + 6.5 + 8 = 25.5$$

Test statistics W

$$W = \min(T^-, T^+)$$
$$= \min(10.5, 25.5) = 10.5$$

Expected value of W

$$u_W = \frac{n \cdot (n+1)}{4} = \frac{8 \cdot (8+1)}{4} = 18$$

Standard deviation

$$\sigma_W = \sqrt{\frac{n \cdot (n+1) \cdot (2 \cdot n+1) - \sum \frac{t_i^3 - t_i}{2}}{24}}$$
$$= \sqrt{\frac{8 \cdot (8+1) \cdot (2 \cdot 8+1) - 15}{24}}$$
$$= 7.1$$

z-value

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{10{,}5 - 18}{7.1} = -1.06$$

Again, noting that you actually need about 20 cases to assume normal distribution of W values.

11.5.6 Effect size in the Wilcoxon signed-rank test

The effect size indicates how large the observed effect is compared to the random noise. There are several measures to calculate the effect size in the Wilcoxon test. A common method is to use *r*, defined as:

$$r = \frac{z}{\sqrt{n}}$$

Where *z* is the standardized test statistic value from the Wilcoxon test and *n* is the total number of observations (i.e., the sum of the sizes of both groups).

The value of r can range from -1 to 1, with values near 0 indicating that there is no effect and values near -1 or 1 indicating a strong effect. The sign of *r* indicates the direction of the effect.

The following table can be used to interpret the effect size (effect size *r* according to Cohen (1988)).

|r| < 0.1 no effect / very small effect
|r| = 0.1 small effect
|r| = 0.3 medium effect
|r| = 0.5 large effect

11.5.7 Example Wilcoxon test with DATAtab

A Wilcoxon test can easily be calculated with DATAtab.

That's how it works with DATAtab:

- Simply copy the table below or your own data into the Statistical Calculator and click on Hypothesis tests.
- Then click on the two variables and select Non-Parametric Test.

Response time in the morning	Response time in the evening
34	45
36	33
41	35
39	43
44	42
37	42
39	43
39	43
45	42

DATAtab will then give you the following results.

If you have more than two dependent variables, you can also easily calculate a Friedman test online. To do this, simply click on more than two metric variables.

Wilcoxon-Test

Summary in words

Hypotheses

Copy Settings

Null hypothesis	Alternative hypothesis
There is no difference between the variables Reaction time morning and Reaction time evening	There is a difference between the variables Reaction time morning and Reaction time evening

Descriptive statistics

Copy Settings

	n	Mean	Median	Standard deviation
Reaction time morning	9	39.33	39	3.57
Reaction time evening	9	40.89	42	4.04

Box plot

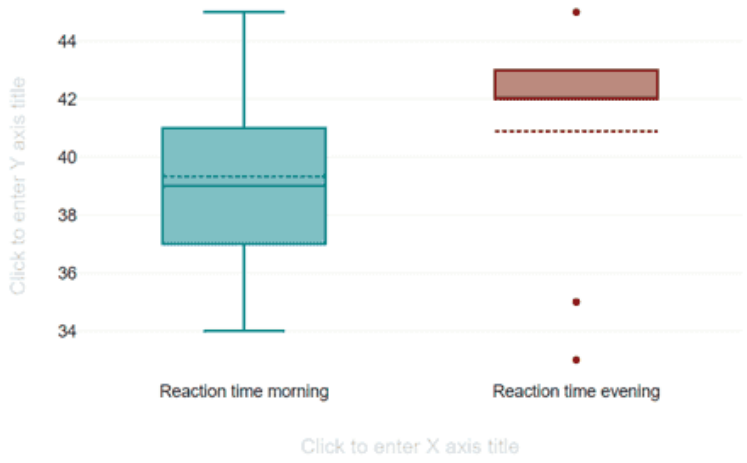
Orientation
- (•) vertical
- ( ) horizontal

Show Points
- ( ) yes
- (•) no

Standard Deviation
- ( ) yes
- (•) no

Size of the graphic
- ( ) small
- (•) medium
- ( ) large
- ( ) extra large

Download png Download svg Settings



Ranks

Copy Settings

		n	Mean Rank	Sum of Ranks
Reaction time evening - Reaction time morning	Negative Ranks	4	3.5	14
	Positive Ranks	5	6.2	31
	Ties	0		
	Total	9		

- Negative Ranks: Reaction time evening < Reaction time morning
- Positive Ranks: Reaction time evening > Reaction time morning
- Ties: Reaction time evening = Reaction time morning

Wilcoxon-Test

Copy Settings

	W	z	p	r
Reaction time evening - Reaction time morning	14	-1.01	.312	0.34

12. Frequency analysis

The following chapter deals with methods for the analysis of frequencies. It starts with a presentation of the binomial test and its execution and interpretation. In the second part, reference is made to the chi-square test and its areas of application.

12.1 Binomial test

The binomial test is a hypothesis test used when there is a categorical variable with two expressions, e.g., gender with male and female. The binomial test can then check whether the frequency distribution of the variable corresponds to an expected distribution, e.g.

- Men and women occur equally often
- The proportion of women is 54%.

This is a special case when it is to be tested in general whether the frequency distribution of the variable has arisen by chance or not. In this case, the probability of occurrence is set to 50%.

The binomial test can therefore be used to check whether the frequency distribution of a sample matches that of the population or not.

12.1.1 Hypotheses in binomial test

The binomial test thus checks whether the **frequency distribution of a variable** with two expressions/categories in the sample corresponds to the distribution in the **population.**

The **hypothesis** in the binomial test results in the **one-tailed case** to:

- **Null hypothesis:** The frequency distribution of the sample corresponds to the distribution of the population.
- **Alternative hypothesis:** The frequency distribution of the sample does not correspond to the distribution of the population.

Thus, the non-directional hypothesis only tests whether there is a difference or not, but not in which direction this difference goes.

In the **two sided case**, the aim is to investigate whether the probability of occurrence of an expression in the sample is greater or less than a given or true percentage.

In this case, an expression is defined as "success" and it is checked whether the true "probability of success" is smaller or larger than that in the sample.

The alternative hypothesis then results in:

- **Alternative hypothesis:** The true probability of success is smaller/larger than the given value.

12.1.2 Binomial test calculation

To calculate a binomial test, you need the sample size, the number of cases that are positive from it, and the probability of occurrence in the population.

Sample size	Successes	Probability
12	3	0.35

Alternative hypothesis	p
True probability of success is less than 0.35	0.347
True probability of success is not equal to 0.35	0.559
True probability of success is greater than 0.35	0.849

12.1.3 Binomial test example

A possible example for a binomial test would be the question whether the gender ratio in the major marketing at the university XY (sample) differs significantly from that of all business students at the university XY (population).

Listed below are the students majoring in marketing; women make up 55% of the total business degree program.

Marketing Student	Gender
1	female
2	male
3	female
4	female
5	female
6	male
7	female
8	male
9	female
10	female

This is how it works with DATAtab:

You can easily recalculate the above example with DATAtab. Insert the upper table including the first row into the statistics calculator.

DATAtab gives you the following result for this sample data:

Statistics

Copy Word Copy Excel

	Category	N	Observed Probability	Expected valid Probability
Gender	female	7	70%	55%
	male	3	30%	
Valid Total	10	100%		

Binomial Test

Copy Word Copy Excel

Alternative hypothesis	p
True probability of success is less than 0.55	.9
True probability of success is not equal to 0.55	.528
True probability of success is greater than 0.55	.266

Each cell in the table of expected observations has five or more observations, so the assumptions for the Chi^2 test are met.

12.1.4 Interpretation of a Binomial Test

With an expected test value of 55%, the p-value is 0.528. This means that the p-value is above the signification level of 5% and the result is therefore not significant. Consequently, the null hypothesis must not be rejected.

In terms of content, this means that the gender ratio of the marketing specialization (=sample) does not differ significantly from that of all business administration students at XY University (=population).

12.2 Chi-square test

The Chi-square test is a hypothesis test that can be used when you want to determine if there is a relationship between two, usually categorical variables.

What are categorical variables? Categorical variables are, for example, a person's gender, preferred newspaper, frequency of watching television or highest level of education. If two categorical variables are to be tested to see whether there is a correlation between them, a Chi-square test is used.

Categorical variables

Gender	Preferred newspaper	Frequency of television	Highest educational level
1 = male	1 = The Washington Post	1 = daily	1 = Without graduation
2 = female	2 = The New York Times	2 = several times per week	2 = College
	3 = USA Today	3 = more rarely	3 = Bachelor's degree
	4 = ...	4 = never	4 = Master's degree

Figure 80: Example of categorical variables

The **chi-square test** is a hypothesis test used with **categorical variables**, i.e., nominal or ordinal scale levels. The chi-square test checks whether the frequencies occurring in the sample differ significantly from those frequencies that would be expected. Thus, the **observed frequencies** are compared with the **expected frequencies** and their deviations are examined.

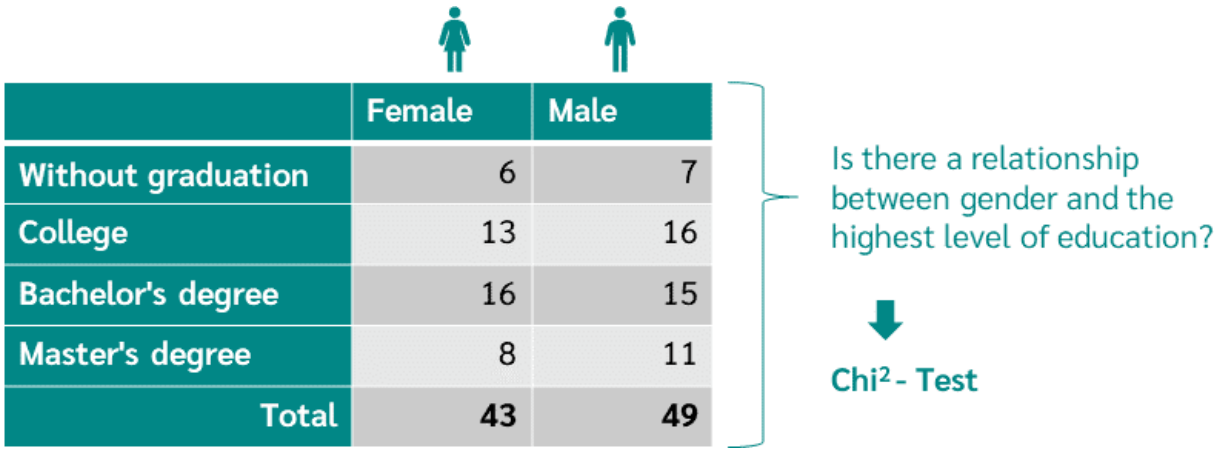


Figure 81: Use of the chi-square test

Let's say we want to investigate whether there is a correlation between gender and the highest educational attainment. To do this, we create a questionnaire in which the participants check off which gender they have and what their highest educational attainment is. The result of the survey is then presented in a cross tabulation.

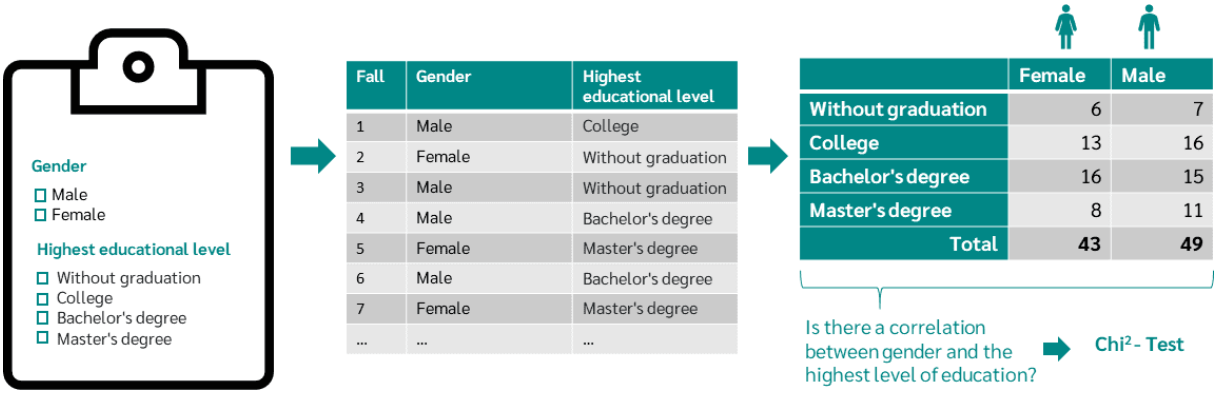


Figure 82: From questionnaire to crosstab

To examine whether there is a relationship between gender and highest educational attainment, the chi-square test is used.

12.2.1 Applications of the Chi-Square Test

There are several applications of the chi-square test, it can be used to answer the following questions:

1) Independence test

Are two categorical variables independent of each other? For example, does gender affect whether or not a person has a Netflix subscription?

2) Distribution test

Are the observed expressions of two categorical variables equal to the expected one? One question could be whether one of the three video streaming services Netflix, Amazon and Disney is subscribed to more frequently than average.

3) Homogeneity test

Do two or more samples come from the same population? One question could be whether the subscription numbers of the three-video streaming services Netflix, Amazon and Disney differ in different age groups.

12.2.2 Calculation of Chi-Square- test

The chi-square value is calculated via:

Observed frequency Expected frequency

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

To clarify the calculation of the chi-square value, we refer to the following case: For **variables one** and **two** with **category A** and **B**, an observation was made, or a sample exists. Now we want to check whether the frequencies from the sample correspond to the expected frequencies from the population.

Observed frequency:

		Variable 2	
		Category A	Category B
Variable 1	Category A	10	13
	Category B	13	14

Expected frequency:

		Variable 2	
		Category A	Category B
Variable 1	Category A	9	11
	Category B	12	13

With the following formula you can now calculate the chi-square:

$$\chi^2 = \frac{(10-9)^2}{9} + \frac{(13-11)^2}{11} + \frac{(13-12)^2}{12} + \frac{(14-13)^2}{13} = 0.635$$

After the chi-square has been calculated, the **degree of freedom df is** still needed. This results from:

$$df = (p-1)(q-1) = 1$$

with

- **p:** number of lines
- **q:** number of columns

From the table of the Chi-square distribution, you can now read off the critical chi-square value. For a significance level of 5%, this is 3.841. Since the calculated chi-square value is smaller, there is no significant difference.

As a **prerequisite for** the chi-square test, it should be noted that all expected frequencies must be greater than five.

12.2.3 Chi-Square Test of Independence

The **chi-square** independence **test** is used when two categorical variables are to be tested for their independence. The aim is to analyze whether the characteristic values of the first variable are influenced by the characteristic values of the second variable and vice versa.

A question in this context could be for example: "Does gender influence whether a person has a Netflix subscription or not? "For the two variables "**gender"** (male, female) and "**possession of a Netflix subscription"** (yes, no), it is checked whether they are independent. If this is not the case, there is a correlation between the characteristics. The research question, which can be answered with the chi-square test, is: Are the characteristics gender and ownership of a Netflix subscription independent of each other?

In order to calculate the Chi-square, an observed and an expected frequency must be given. In the independence test, the expected frequency is the frequency that results when both variables are independent. If two variables are independent, the expected frequencies of the individual cells result with

$$f(i,j) = \frac{\text{RowSum}(i) \cdot \text{ColumnSum}(j)}{N},$$

where i and j are the number of rows and columns in the table, respectively.

For the fictitious Netflix example, the following tables could result. On the left is the table with the frequencies observed in the sample and on the right the table that would result if perfect independence existed.

Observed frequency:

	Male	Female
Netflix Yes	10	13
Netflix No	15	14

Expected frequency if independent:

	Male	Female
Netflix Yes	(23 · 25) / 52 = 11.06	(23 · 27) / 52 = 11.94
Netflix No	(29 · 25) / 52 = 13.94	(29 · 27) / 52 = 15.06

The Chi-square is then calculated as:

$$\chi^2 = \frac{(10-11,06)^2}{11,06} + \frac{(13-11,94)^2}{11,94} + \frac{(15-13,94)^2}{13,94} + \frac{(14-15,06)^2}{15,06} = 0.35$$

.

From the Chi-square table you can now read the critical value again and compare it with the result.

The **assumptions for** the **Chi-square independence test** are that the observations are from a random sample and that the expected frequencies per cell are greater than 5.

12.2.4 Chi-square distribution test

If a variable is present with two or more expressions, the differences in the frequency of the individual expressions can be examined.

The **Chi-square distribution test**, or **goodness-of-fit test**, checks whether the frequencies of the individual characteristic values in the sample correspond to the frequencies of a defined distribution. In most cases, this defined distribution is that of the population. In this case, it is tested whether the sample originates from the respective population.

Here is an example:

For market researchers, it could be of interest in this difference whether there is a difference in the market penetration of the three-video streaming services Netflix, Amazon and Disney between Berlin and all of Germany. The expected frequency is then the distribution of streaming services across Germany and the observed frequency results from a survey in Berlin. The following tables show the fictitious results for the chi-square goodness-of-fit test:

Observed frequency in Berlin:

Video Service	Frequency
Netflix	25
Amazon	29
Disney	13
Others or none	20

Expected frequency (all Germany):

Video Service	Frequency
Netflix	23
Amazon	26
Disney	16
Other or none	22

The Chi-square then results in

$$\chi^2 = \frac{(25-23)^2}{23} + \frac{(29-26)^2}{26} + \frac{(13-16)^2}{16} + \frac{(20-22)^2}{22} = 2.389$$

12.2.5 Chi-square homogeneity test

The Chi-square homogeneity test can be used to check whether two or more samples come from the same population? One question could be whether the subscription numbers of the three-video streaming services Netflix, Amazon and Disney differ in different age groups. As a fictitious example, a survey is conducted in three age groups with the following result:

Observed frequency:

Age in years	15-25	25-35	35-45
Netflix	25	23	20
Amazon	29	30	33
Disney	11	13	12
Others or none	16	24	26

As with the Chi-square independence test, this result is compared with the table that would result if the distributions of Streaming providers were independent of age.

12.2.6 Effect size for Chi-square test

So far, we only know whether we can reject the null hypothesis or not, but it is very often also of great interest to know how strong the correlation of the two variables is. This can be answered with the help of the effect size.

In the chi-square test, **Cramer's V** can be used to calculate the effect size. Here, a value of 0.1 is considered small, one of 0.3 medium, and a value of 0.5 large.

Effect size	Cramér's V
Small	0.1
Medium	0.3
Large	0.5

12.2.7 Effect size vs. p-value

Please note that the p-value does not tell you anything about the **strength of the correlation or the effect** and the p-value depends on the sample size!

The following points should therefore be considered:

- If there is a correlation in the population, the larger the sample, the more clearly it is indicated in the p-value.
- If the sample is very large, very small correlations can be detected in the population.
- These small contexts may not even be relevant anymore.

Therefore, if there is in a small sample and in a large sample an equal effect, the p-values would still differ.

The larger the sample, the smaller the p-value, and thus **very small associations** can be confirmed with a **very large sample.**

This is where the **effect size** becomes important. With the effect size in the chi-square test, differences across several studies can be made comparable.

12.2.8 Example: Chi-square test with DATAtab

Independence test

As an example of a chi-square test examining independence, consider the use of umbrellas. On a rainy day, it was counted how many women and how many men come to the university with an umbrella.

The results of this count are listed in the table below:

Gender	Using an Umbrella
female	Yes
male	Yes
female	Yes
female	Yes
male	Yes
male	No
female	No
male	No
female	No
female	No
male	No
female	Yes
male	Yes
female	Yes
male	Yes
male	Yes
male	No
female	No
male	No
female	No
female	No

Gender	Using an Umbrella
female	No

The research question in this context is "Is the difference in the use of an umbrella between women and men statistically significant or random?"

This is how it works with DATAtab:

After copying the above table into the hypothesis testing calculator, you can calculate the **chi-square test.**

To do this, simply click on the two variables "*gender*" and "*umbrella included*". As a result, you get the (1) cross-tabulation, the (2) expected frequency with perfectly independent variables and the (3) chi-square test.

		Umbrella included		
		yes	no	Total
Gender	female	5	7	12
	male	5	5	10
	Total	10	12	22

Expected frequencies for perfectly independent variables:

		Umbrella included		
		yes	no	Total
Gender	female	5.455	6.545	12
	male	4.545	5.455	10
	Total	10	12	22

Chi-squared test	
Chi-squared	0.153
df	1
p value	0.696

With an α-level of 5% and a degree of freedom of one, the **table of chi-square values** yields a critical value of 3.841. Since the calculated chi-square value is smaller than the critical value, there is no significant difference in this example. The null hypothesis is thus retained. In terms of content, this means that men and women do not differ regarding the frequency of their screen use.

Distribution test

In a Viennese district, the party affiliation of 22 people was surveyed. Now it is to be examined whether the residents of the district (sample) have the same voting behavior as the residents of the entire city of Vienna (population). It is known, that Party A has a share of 40% in Vienna and Party C has a share of 35%.

Party
Party A
Party C
Party A
Party C
Party A
Party C
Party B
Party B
Party C
Party A
Party C
Party A
Party A
Party B
Party B
Party A

Party
Party A
Party B
Party A
Party A
Party C
Party C

To calculate the chi-squared test for the example, simply copy the upper table into the Hypothesis Test Calculator.

Party A has a 40% share in Vienna and *party C* has 35%. You will therefore now receive the following results:

	Category	N	Observed Probability	Expected Probability
Party	Party A	10	45.455%	40%
	Party C	7	31.818%	35%
	Party B	5	22.727%	
	Total	22	100%	

Chi-squared test	
Chi-squared	0.264
df	2
p	0.876

If the significance level is set at 0.05, the p-value calculated at 0.876 is greater than the significance level. Thus, the null hypothesis is not rejected and it can be assumed that the residents of the district have the same voting behavior as the residents of the entire city of Vienna.

13. Statistical tests to test for differences in more than two groups.

In the next chapter, we deal with statistical tests for examining differences when more than two groups are present. In this case, methods of analysis of variance are used. A distinction must be made between one-factor and two-factor analysis of variance and between analyses of variance with and without repeated measures.

13.1 Analysis of Variance (ANOVA)

An analysis of variance (ANOVA) tests whether statistically significant differences exist between more than two samples. For this purpose, the means and variances of the respective groups are compared with each other. In contrast to the t-test, which tests whether there is a difference between two samples, the ANOVA tests whether there is a difference between more than two groups.

There are different types of analysis of variance, being the one-way and two-way analyses of variance the most common ones, each of which can be calculated either with or without repeated measurements.

In this tutorial you will learn the basics of ANOVA; for each of the four types of analysis of variance you will find a separate detailed tutorial:

- One-factor (or one-way) ANOVA
- Two-factors (or two-way) ANOVA
- One-factor ANOVA with repeated measurements
- Two-factors ANOVA with repeated measurements

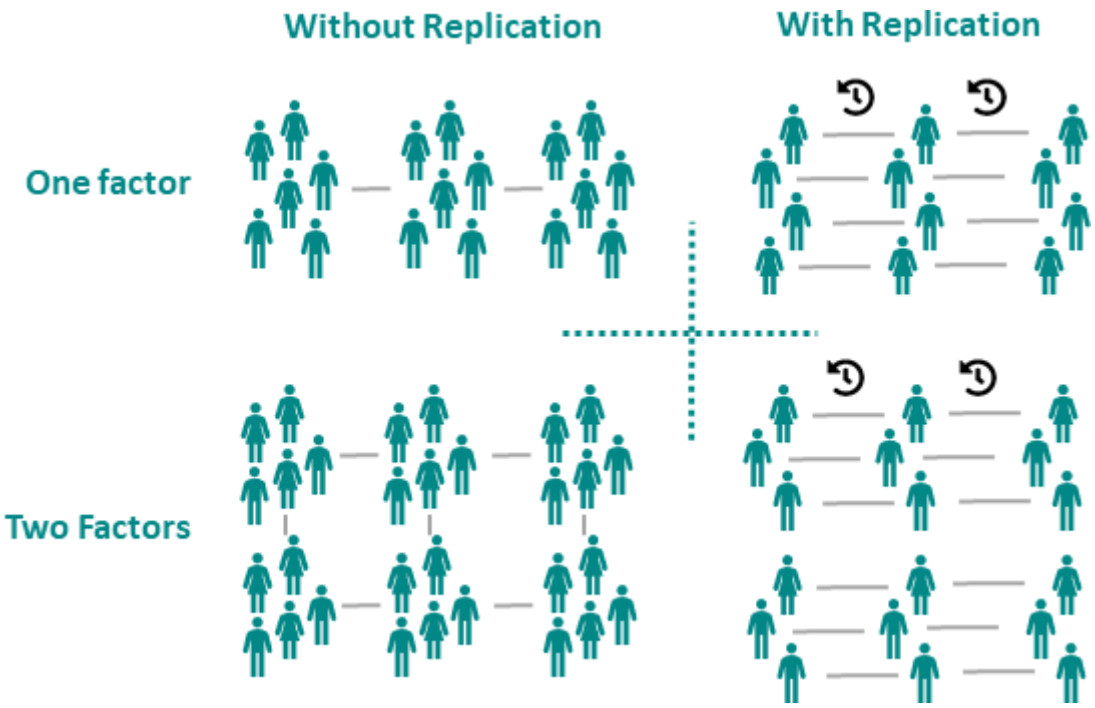


Figure 83: Types of analysis of variance

13.1.1 Why not calculate multiple t-tests?

ANOVA is used when there are more than two groups. Of course, it would also be a possibility to calculate a t-test for each combination of the groups. The problem here, however, is that every hypothesis test has some degree of error. This probability of error is usually set at 5%, so that, from a purely statistical point of view, every 20th test gives a wrong result

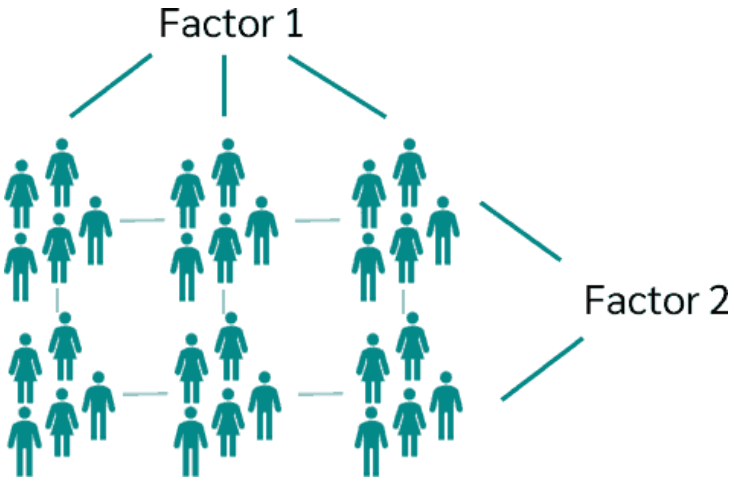
If, for example, 20 groups are compared in which there is actually no difference, one of the tests will show a significant difference purely due to the sampling.

13.1.2 Difference between one-way and two-way ANOVA

The one-way analysis of variance only checks whether an independent variable has an influence on a metric dependent variable. This is the case, for example, if it is to be examined whether the place of residence (independent variable) has an influence on the salary (dependent variable). However, if two

factors, i.e. two independent variables, are considered, a two-way analysis of variance must be used.

One-factor ANOVA	Two-factors ANOVA
Does a person's place of residence (independent variable) influence his or her salary?	Does a person's place of residence (1st independent variable) and gender (2nd independent variable) affect his or her salary?



13.1.3 Analysis of variance with and without repeated measures

Depending on whether the sample is independent or dependent, either analysis of variance with or without repeated measures is used. If the same person was interviewed at several points in time, the sample is a dependent sample and analysis of variance with repeated measurements is used.

13.2 One-factor ANOVA

The one-way analysis of variance is an extension of the t-test for independent groups. With the t-test only a maximum of two groups can be compared; this is now extended to more than two groups. For two groups (k = 2), the analysis of variance is therefore equivalent to the t-test. The independent variable is accordingly a nominally scaled variable with at least two characteristic values. The dependent variable is on a metric scale. In the case of the analysis of variance, the independent variable is referred to as the factor.

The following question can be answered: Is there a difference in the population between the different groups of the independent variable with respect to the dependent variable?

The aim of ANOVA is to explain as much variance as possible in the dependent variable by dividing it into the groups. Let us consider the following example.

13.3 One-factor ANOVA example

With the help of the dependent variable, e.g. "highest educational qualification" with the three characteristics group 1, group 2 and group 3 should be explained as much variance of the dependent variable "salary" as possible. In the graphic below, under A) a lot of variance can be explained with the three groups and under B) only very little variance.

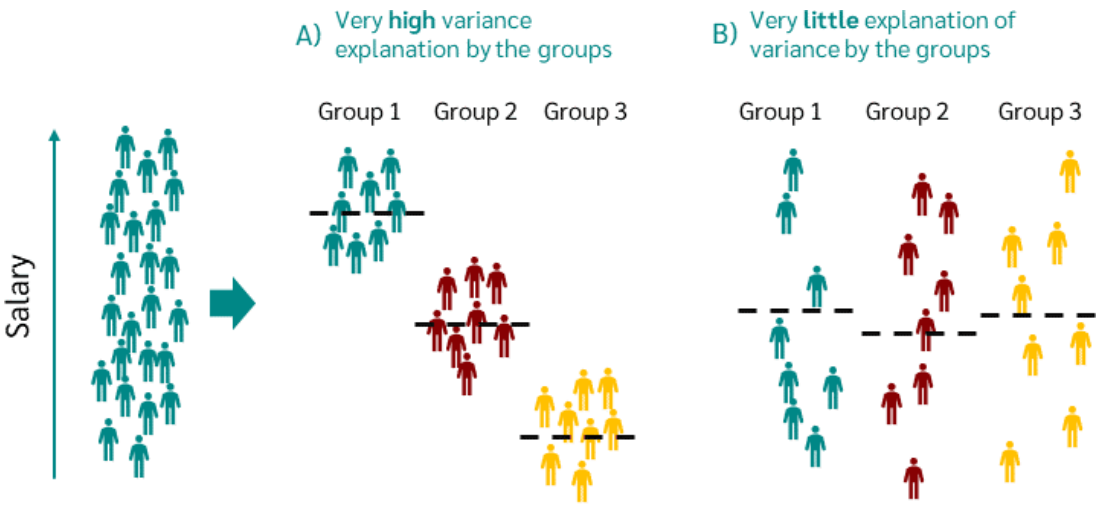


Figure 84 Variance elucidation of the ANOVA

Accordingly, in case A) the groups have a very high influence on the salary and in case B) they do not.

In the case of A), the values in the respective groups deviate only slightly from the group mean, the variance within the groups is therefore very small. In the case of B), however, the variance within the groups is large. The variance between the groups is the other way round; it is large in the case of A) and small in the case of B). In the case of B) the group means are close together, in the case of A) they are not.

	Variance within the groups	Variance between group means
Case A)	Small	Large
Case B)	Large	Small

13.4 Analysis of variance hypotheses

As with the statistical tests already discussed, when performing analyses of variance, it is necessary to formulate hypotheses in advance that are to be tested. The null hypothesis and the alternative hypothesis arise in a single factor analysis of variance as follows:

- Null hypothesis H_0: The mean of all groups is equal.
- Alternative hypothesis H_1: There are differences in the means of the groups.

The results of the Anova can only make a statement about whether there are differences between at least two groups. However, it cannot be determined which groups are exactly different. A post-hoc test is needed to determine which groups differ. There are various methods to choose from, with Duncan, Dunnet C and Scheffe being among the most common methods.

Example:

In a screw factory, a screw is produced by three different production lines. You now want to find out whether all production lines produce screws with the same weight. To do this, take 50 screws from each production line and measure the weight. Now you use the ANOVA procedure to determine whether the average weight of the screws from the three production lines differs significantly from one another.

An example of the one-way analysis of variance would be to investigate whether the daily coffee consumption of students from different fields of study differs significantly.

	Dependent variable	**Independent variable**
Level of measurement	An metric-scaled variable	A nominally scaled variable with more than two levels
Example	Weekly coffee consumption	Subject (math, psychology, economics)

13.5 Assumptions for one-way analysis of variance

Before you start performing an analysis of variance, it is important to check the following assumptions so that you know whether your data are suitable for this test. These assumptions are:

1) **Scale level:** The scale level of the dependent variable must be metric, whereas the independent variable must be nominally scaled.
2) **Homogeneity:** The variances in each group should be roughly the same. This can be checked with the Levene test.
3) **Normal distribution:** The data within the groups should be normally distributed. This means that the majority of the values are in the average range, while very few values are significantly below or significantly above. If this condition is not met, the Kruskal-Wallis test can be used.

If there are no independent samples but dependent ones, then a one-factor analysis of variance with repeated measures is used.

13.6 Welch's ANOVA

If the condition of variance homogeneity is not fulfilled, **Welch's** ANOVA can be calculated instead of the "normal" **ANOVA.** If the Levene test results in a significant deviation of the variances in the groups, DATAtab automatically calculates the Welch's ANOVA in addition.

Welch's ANOVA

Copy Word Copy Excel

	F	df1	df2	p-value
Welch-Test	1.9	2	5.34	0.238

13.7 Effect size Eta squared (η^2)

The best known measures of effect size for analysis of variance are the Eta squared and the partial Eta squared. For a single factor ANOVA, the Eta squared and the partial Eta squared are identical.

The Eta squared estimate the variance that a variable explains. however, it should be noted that the variance explained is always overestimated. Eta squared is calculated by dividing the sum of squares between by the sum of squares total.

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

13.8 Two factor analysis of variance

As the name suggests, two-way analysis of variance examines the influence of two factors on a dependent variable. This extends the one-way analysis of variance by a further factor, i.e. by a further nominally scaled independent

variable. The question is again whether the mean of the groups differs significantly.

	Dependent variable	Independent variable
Level of measurement	One metric-scaled variable	Two nominally scaled variables
Example	Weekly coffee consumption	Subject (math, psychology, economics) and semester (winter, summer)

Example:

In a screw factory, a screw is produced by three different production systems, factor 1 in two shifts, factor 2. You now want to find out whether the production facilities or the shifts have an influence on the weight of the bolts. To do this, take 50 screws from each production line and each shift and measure the weight. Now you use two-factor ANOVA to determine whether the average weight of the screws from the three production lines and the two shifts is significantly different from one another.

13.9 Calculate example with DATAtab

One-way analysis of variance:

You want to check whether there is a difference in coffee consumption between students in different subjects. To do this, ask 10 students from each field of study.

Coffee consumption	Subject
21	Math
23	Math
18	Economics
22	Economics
...	...

This is how it works with DATAtab:

- First you copy the table above into the statitics calculator,
- click on Hypothesis test and
- select the three variables.

The result looks like this:

One-way analysis of variance:

One-way analysis of variance:

	n	Mean	SD
Math	10	16.6	7.291
Economics	10	19.8	4.131
Psychology	10	17.8	6.443
Total	30	18.067	5.938

	Sum of squares	df	Mean of squares	F	p
Between the groups	52.267	2	26.133	0.702	0.505
Within the groups	1005.6	27	37.244		
Total	1057.867	29			

Where N is the number of cases for each category, df is the degrees of freedom, F is the F-statistic from the calculated analysis of variance and p is the p-value.

13.10 Repeated Measures ANOVA

Repeated measures ANOVA tests whether there are statistically significant differences in three or more dependent samples. In a dependent sample, the same participants are measured multiple times under different conditions or at different time points.

The one-way analysis of variance with repeated measures is the extension of the t-test for dependent samples for more than two groups.

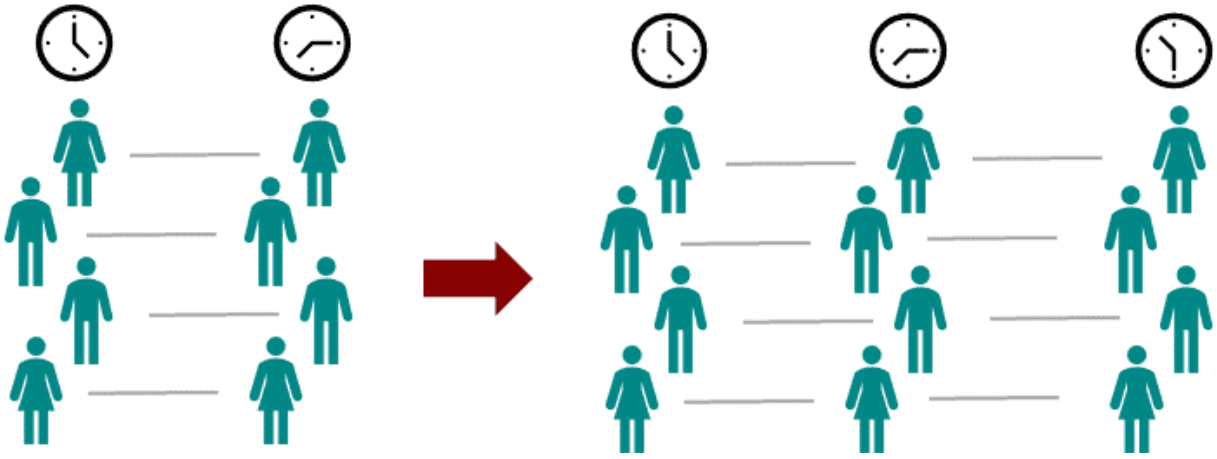


Figure 85: Measurement repetitions

13.10.1 What are dependent samples?

In a dependent sample, the measured values are connected. For example, if a sample is drawn of people who have knee surgery and these people are interviewed before the surgery and one week and two weeks after the surgery, it is a dependent sample. This is the case because the same person was interviewed at two points in time.

Repeated measures: Measurements are repeated when a person is questioned at different times. This is the case, for example, when a person is asked about the intensity of the pain after 3, 6 and 9 months after a surgery.

Now, of course, it doesn't have to be about people or points in time, in a generalized way, we can say: In a dependent sample, the same test units are

measured several times under different conditions. The test units can be people, animals or cells, for example, and the conditions can be time points or treatments, for example.

13.10.2 Difference of analysis of variance with and without repeated measurements

If 3 or more independent samples are available, ANOVA without repeated measures is used. But be careful, of course the assumptions have to be checked.

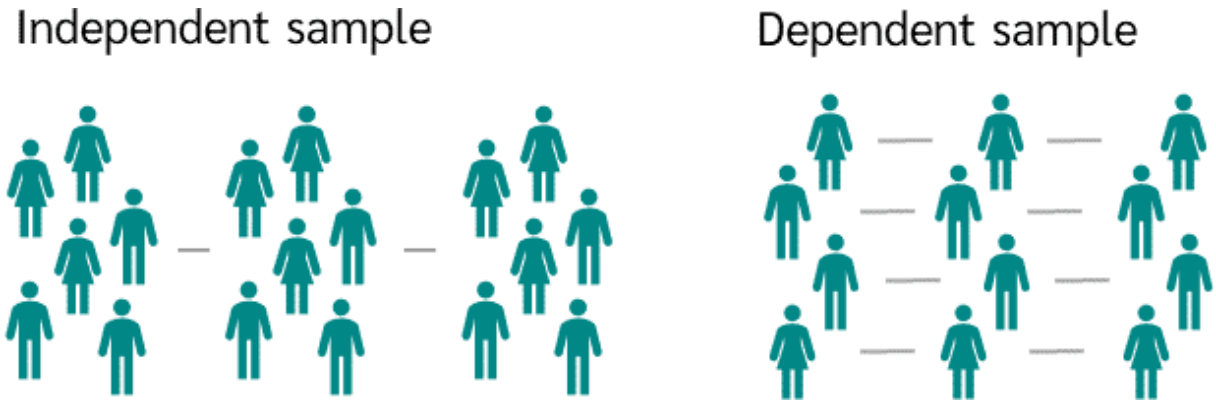


Figure 86: Independent and dependent sample

13.10.3 Example of repeated measures ANOVA

You might be interested to know whether therapy after a slipped disc has an influence on the patient's perception of pain. For this purpose, you measure the pain perception before the therapy, in the middle of the therapy and at the end of the therapy. Now you want to know if there is a difference between the different times.

So, your independent variable is time, or therapy progressing over time. Your dependent variable is the pain perception. You now have a history of the pain perception of each person over time and want to know whether the therapy has an influence on the pain perception.

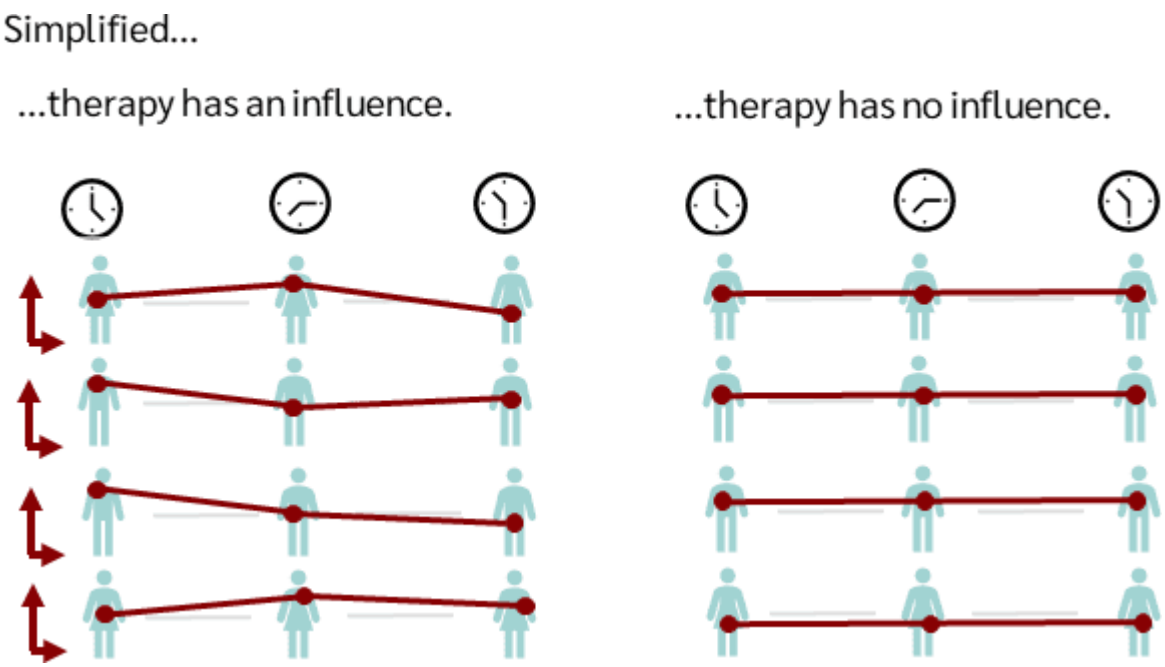


Figure 87: Example independet and dependet sample

To put it simply, in the left case the therapy has an influence and in the right case the therapy has no influence on the pain sensation. In the course of time, the pain sensation does not change on the right hand case, but it does on the left hand one.

13.10.4 Research question and hypotheses

What is the research question in a repeated measures ANOVA? The research question is: Is there a significant difference between the dependent groups in terms of the mean?

The null and alternative hypothis results in:

- **Null hypothesis:** there are no significant differences between the dependent groups.
- **Alternative hypothesis:** there is a significant difference between the dependent groups.

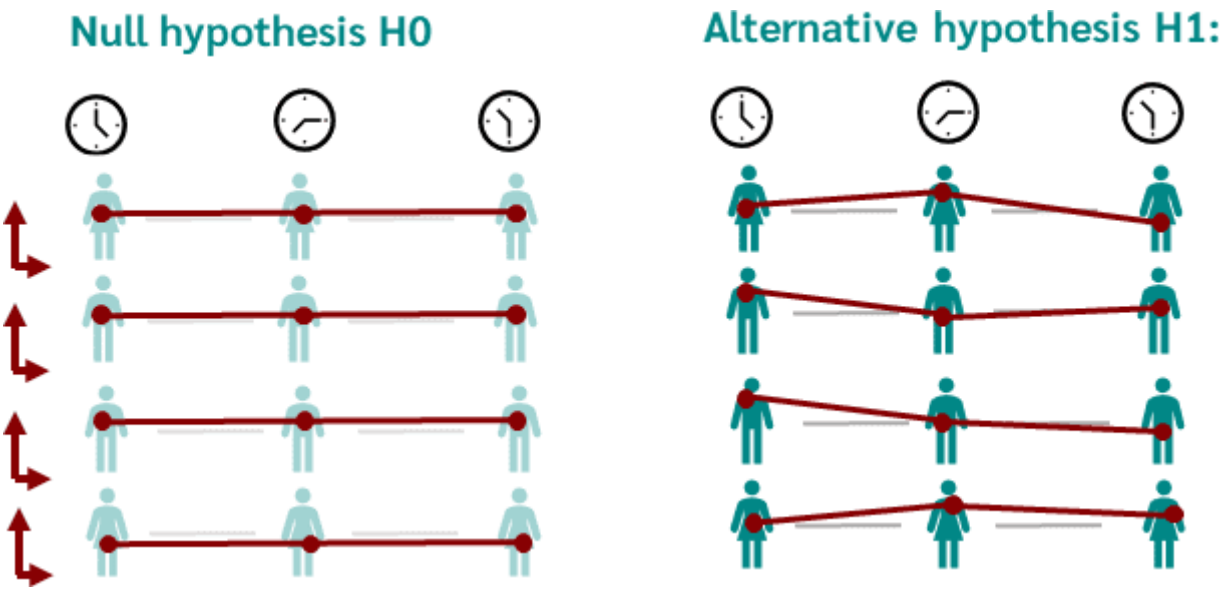


Figure 88: Example null and alternative hypothesis

13.10.5 Assumptions ANOVA with repeated measures

Now we come to the assumptions of ANOVA with repeated measures and finally I will show you how you can easily calculate it online. So what are the assumptions?

- **Dependent samples:** The samples must be dependent samples.
- **Normality:** The data should be approximately normally distributed and have metric scale level. This assumption is especially important when the sample size is small. When the sample size is large, ANOVA is somewhat robust to violations of normality.
- **Sphericity:** The variances of the differences between all combinations of factor levels (time points) should be the same.
- **Homogeneity of Variances:** The variance in each group should be equal. Levene's test can be used to check this assumption.
- **Homogeneity of Covariances (Sphericity):** The variances of the differences between all combinations of the different groups should be equal. This assumption can be tested using Mauchly's test of sphericity.
- **No significant Outliers:** Outliers can have a disproportionate effect on ANOVA, potentially leading to misleading results.

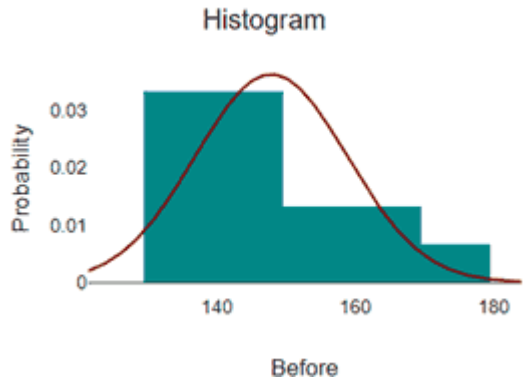
Whether data is normally distributed or not can be tested using the QQ plot or the Kolmogorov smirnov test.

Tests for normal distribution of Before

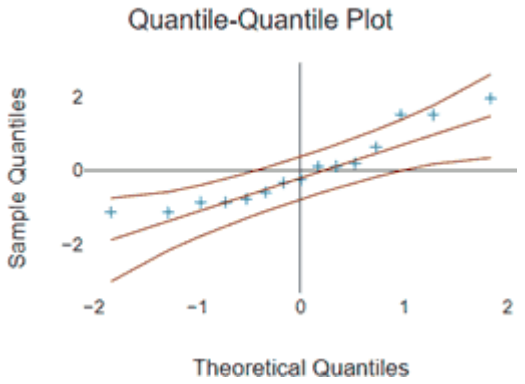
Copy AI Interpretation

	Statistics	p
Kolmogorov-Smirnov	0.16	.788
Kolmogorov-Smirnov (Lilliefors Corr.)	0.16	.407
Shapiro-Wilk	0.9	.085
Anderson-Darling	0.61	.113

Download png Download svg



Download png Download svg



Whether the assumption of sphericity is violated can be tested using Mauchly's test for sphericity. If the resulting p-value is greater than 0.05, it can be assumed that the variances are equal, and the condition is not violated.

Mauchly's Test of Sphericity

Copy

Mauchly's W	Chi-Square	df	p	Greenhouse-Geisser ε	Huynh-Feldt ε
0.85	2.06	2	.357	0.87	0.99

If this assumption is violated, adjustments such as Greenhouse-Geisser or Huynh-Feldt can be made.

13.10.6 Results of the one-factor analysis of variance with repeated measures.

The analysis of variance with repeated measurement gives you a p-value for your data. With the help of this p-value you can read whether there is a significant difference between the repeated measurements.

Repeated Measures ANOVA

Copy AI Interpretation

	Type III Sum of Squares	df	Mean Square	F	p	η^2
Treatment	524.93	2	262.47	5.44	.01	0.28
Error	1351.73	28	48.28			

If the calculated p-value is smaller than the predefined significance level, which is usually 0.05, the null hypothesis is rejected.

In this example, the p-value is 0.01, which is less than 0.05. Therefore, the null hypothesis is rejected and it can be assumed that there is a difference between the different time points.

13.10.7 Effect size for repeated measures ANOVA

In the case of analysis of variance with repeated measures, the effect size can be calculated via the partial eta squared (η^2_p). Here, the variance within individuals is related to the variance that cannot be explained, i.e. the error variance.

$$\eta_p^2 = \frac{SS_{Treatment}}{SS_{Treatment} + SS_{Error}}$$

13.10.8 Bonferroni Post-hoc-Test

As soon as there is a significant difference between the different time points, it is of course also of interest to identify between which exact time points that difference exists. This can be found out with the help of the Bonferroni post-hoc test.

In the Bonferroni post-hoc test in a repeated measures ANOVA, multiple t-tests are calculated for dependent samples. However, the problem with multiple testing is that the so-called alpha error (the false rejection of the null hypothesis) increases with the number of tests. To counteract this, the Bonferroni post-hoc test calculates the obtained p-values times the number of tests.

Bonferroni Post-hoc-Tests

Copy

		Mean diff.	Std. Error	t	p	95% CI lower limit	95% CI upper limit
Before	Middle	0.93	2.792	0.334	1	-5.06	6.92
Before	End	7.67	2.746	2.792	.043	1.78	13.56
Middle	End	6.73	1.994	3.377	.014	2.46	11.01

In the present case, 3 tests were performed, so for the calculation of the Bonferroni post-hoc test, the p-value obtained from the t-test was multiplied by 3 in the background. If one or more p-values are less than 0.05, a significant difference between the two groups is assumed. In this case, we therefore have a significant difference between Before and End and between Middle and End.

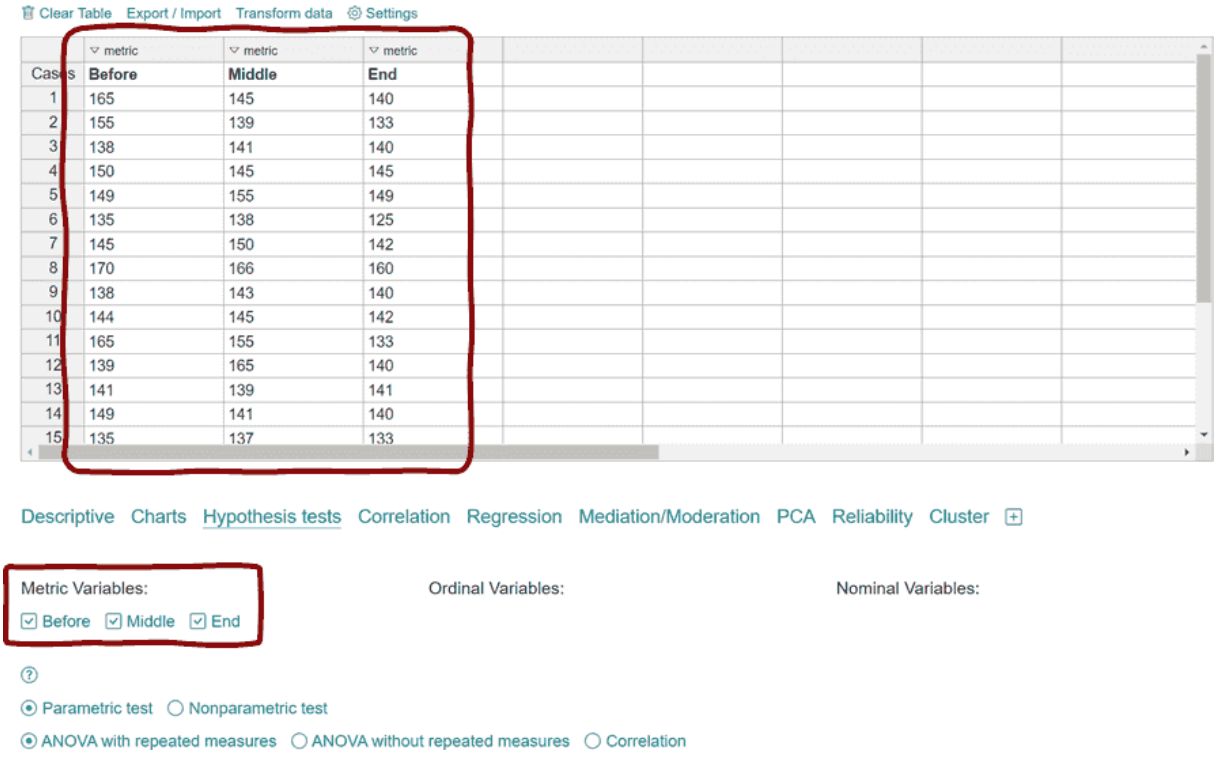
13.10.9 Calculate ANOVA with measurement repetitions with DATAtab

ANOVA with repeated measures can be easily calculated with DATAtab. To do this, simply visit the repeated measures ANOVA calculator on DATAtab and copy your own data into the table.

Now you just need to select your variables. If you select three or more metric variables, an analysis of variance with repeated measures is automatically calculated.

And you get the results. You can read the p value in the table and if you don't know exactly how to interpret the results, just look at the interpretation in words.

In addition, the results are displayed in a boxplot. Finally, the Bonferroni post toc test is calculated.



13.10.10 Calculate a repeated measures ANOVA by hand

How do you calculate an analysis of variance with repeated measures by hand? Here you can find the formulas to calculate an ANOVA.

Let's say this is our data. We have 8 people, each of whom we measured at three different points in time (start, middle and end).

Case	Start	Middle	End	Mean
1	7	9	8	8
2	2	3	3	2.7
3	5	7	6	6
4	6	6	4	5.3
5	4	7	5	5.3
6	7	8	4	6.3
7	4	6	4	4.7
8	5	3	7	5
Mean	5	6.1	5.1	5.4

Mean value of all data

$$G = \frac{\sum x}{N}$$

$$G = \frac{7+9+\ldots+3+7}{24} = 5.4$$

Mean of the time points

$$A_i = \frac{\sum^{Group} x}{n_{Group}}$$

$$A_1 = \frac{7+2+\ldots+4+5}{8} = 5$$

$$A_2 = \frac{9+3+\ldots+6+3}{8} = 6,1$$

$$A_3 = \frac{8+3+\ldots+4+7}{8} = 5.1$$

Mean value of cases

$$P_i = \frac{\sum^{VP} x}{p}$$

$$P_1 = \frac{7+9+8}{3} = 8$$

$$\ldots = \ldots$$

$$P_8 = \frac{5+3+7}{3} = 5$$

Sum of squares within subjects

$$QS_{in} = \sum^{Groups}\sum^{VP}(x_{mi} - P_i)^2$$

$$QS_{in} = (7-8)^2 + (9-8)^2 + \ldots + \ldots + (3-5)^2 + (7-5)^2 + \ldots = 31.3$$

Square of Sum Treatmeant

$$QS_{treat} = n \sum^{Groups} (A_i - G)^2$$

$$QS_{treat} = 8 \cdot [(5-5.4)^2 + (6.1-5.4)^2 + (5.1-5.4)^2]$$

$$QS_{treat} = 6.1$$

Square of Sum Res

$$QS_{Res} = \sum^{Groups}\sum^{VP}(x_{mi} - A_i - P_m + G)^2$$

$$QS_{Res} = (5-5)^2 + (3-5)^2 + \ldots + \ldots + (7-6)^2 + (6-6)^2 + \ldots = 25.3$$

Mean Square

$$MQ_{treat} = \frac{QS_{treat}}{df_{treat}} = \frac{6.1}{2} = 3.04$$

$$MQ_{res} = \frac{QS_{res}}{df_{res}} = \frac{25.3}{14} = 1.8$$

F-Value

$$F = \frac{MQ_{treat}}{MQ_{res}} = \frac{3.04}{1.8} = 1.69$$

p-Value

$$p(F, df_{treat}, df_{res}) = 0.22$$

First, we can calculate the necessary mean values. With the mean values we can calculate the Sum of squares and the Mean Square.

Now we can calculate the F value, which is calculated by dividing the mean square of the treatment by the mean square of the residual or error.

Finally we can calculate the p value using the F value and the degrees of freedom from the treatment and error. To calculate the p-Value we use the F distribution.

Get Started

First steps with DATAtab
Statistics Playbooks

Statistics

Descriptive and Inferential Statistics
Level of Measurement

Location parameter
Dispersion parameter
Frequency table
Contingency table
Charts
Bar Chart
Box Plot
Bland-Altman Plot

Hypothesis
Hypothesis Testing
p-Value

Test of Normality
Dependent and independent samples
Confidence interval

F distribution table

The F distribution is used, for example, in the interpretation of an ANOVA. The F-distribution results from the quotient of two chi-square distributions which are divided by the respective degrees of freedom.

Here you can either calculate the critical F-value or the p-value with given degrees of freedom. You can also read the critical F-value for a given alpha level in the tables below.

F critical		alpha	df numerator	df denominator
4.459	=	0.05	2	8

p-Value		F-Value	df numerator	df denominator
0.22	=	1.69	2	14

Table for F distribution 0.95 (a=0.05)

The following table shows the inverse distribution function of the F-distribution for (1-a) = 0.95. On the axes you will find the degrees of freedom of the numerator and denominator.

df2\df1	1	2	3	4	5	6	7	8	9	10
1	161.448	199.5	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19	19.164	19.247	19.296	19.33	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786

13.11 Two-way ANOVA (without repeated measures)

As the name suggests, the **two-way or two-factor analysis** of variance examines the influence of **two factors** on a dependent variable. Here, the single-factor analysis of variance is extended by a further factor, i.e. by a further nominally scaled independent variable. The question here is again whether the mean values of the groups differ significantly.

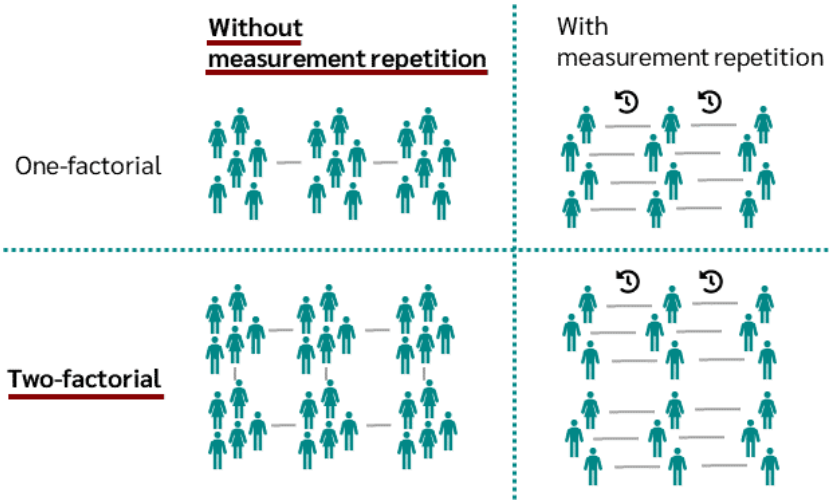
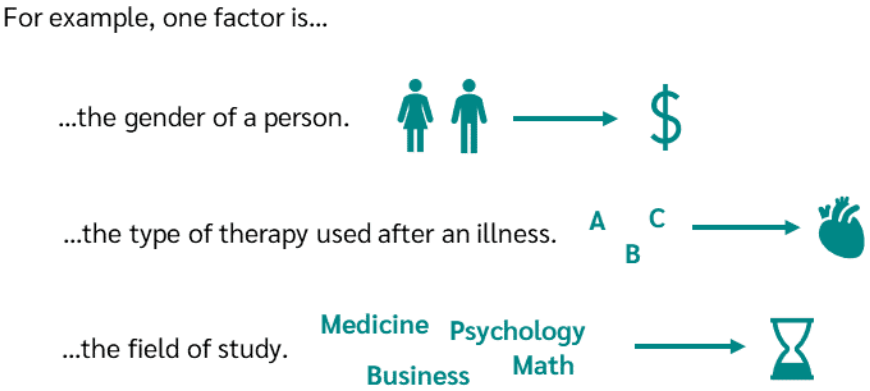


Figure 89: Two-way ANOVA

13.11.1 What is a factor?

A factor is, for example, the gender of a person with the characteristics male and female, the form of therapy used for a disease with therapy A, B and C or the field of study with, for example, medicine, business administration, psychology and math.



In the case of variance analysis, a factor is a categorical variable. You use an analysis of variance whenever you want to test whether these categories have an influence on the so-called dependent variable.

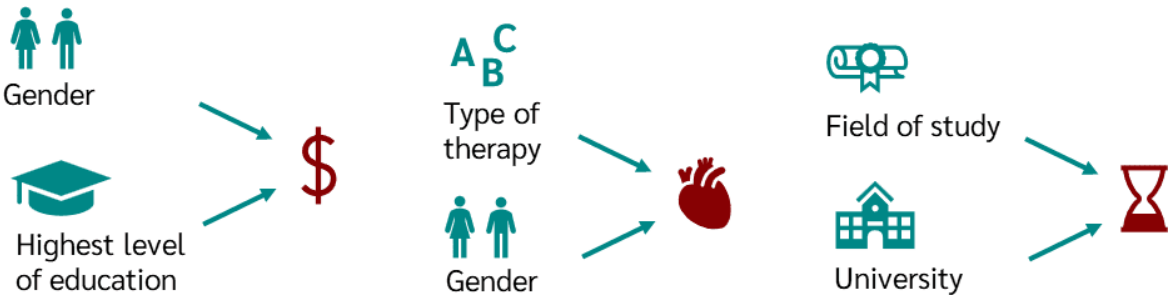
For example, you could test whether gender has an influence on salary, whether therapy has an influence on blood pressure, or whether the field of study has an influence on the duration of studies. Salary, blood pressure or study duration are then the dependent variables. In all these cases you now check whether the factor has an influence on the dependent variable.

Since you only have one factor in these cases, you would use a single factor analysis of variance in these cases (except of course for the gender, there we have a variable with only two expressions, there we would use the t-test for independent samples).

13.11.2 Two factors

Now you may have another categorical variable that you want to include as well. You might be interested in whether:

- in addition to gender, the highest level of education also has an influence on salary.
- besides therapy, gender also has an influence on blood pressure.
- in addition to the field of study, the university attended also has an influence on the duration of studies



Now in all three cases you would not have one factor, but two factors each. And since you now have two factors, you use the two-way analysis of variance.

Using the two-way analysis of variance, you can now answer three things:

- Does factor 1 have an effect on the dependent variable?
- Does factor 2 have an effect on the dependent variable?
- Is there an interaction between factor 1 and factor 2?

Therefore, in the case of one-way analysis of variance, we have one factor from which we create the groups. In the case of two-way analysis of variance, the groups result from the combination of the expressions of the two factors.

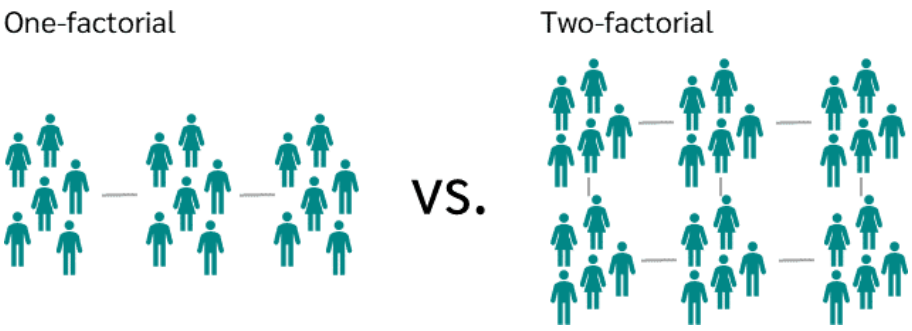


Figure 90: One-factorial vs. two-factorial ANOVA

13.11.3 Example Two-Way ANOVA

Here's an example dataset for a two-way ANOVA in medicine. Let's say we are interested in studying the effect of two factors, "Treatment" and "Gender," on the response variable "Blood Pressure."

In this example, we have two levels of the "Treatment" factor (A and B) and two levels of the "Gender" factor (Male and Female). The "Blood Pressure" measurements are recorded for each participant based on their treatment and gender.

To perform a two-way ANOVA on this dataset, we would test the null hypothesis that there is no interaction between the "Treatment" and "Gender" factors and no main effects of each factor on the "Blood Pressure" response variable.

13.11.4 Hypotheses

Three statements can be tested with the 2 way ANOVA, so there are 3 null hypotheses and therefore 3 alternative hypotheses.

Null hypotheses H_0	**Alternative hypotheses H_1**
There are **no** significant differences in the mean between the groups (factor levels) of the first factor.	There is a significant difference in the mean between the groups (factor levels) of the first factor.
There are **no** significant differences in the mean between the groups (factor levels) of the second factor.	There is a significant difference in the mean between the groups (factor levels) of the second factor.
One factor has **no** effect on the effect of the other factor.	One factor has an effect on the effect of the other factor.

13.11.5 Assumptions

For a two-way analysis of variance to be calculated without repeated measures, the following assumptions must be met:

- The scale level of the dependent variable should be metric, that of the independent variable (factors) nominal scale.
- Independence: The measurements should be independent, i.e. the measured value of one group should not be influenced by the measured value of another group. If this were the case, we would need an analysis of variance with repeated measures.
- Homogeneity: The variances in each group should be approximately equal. This can be checked with Levene's test.
- Normal distribution: The data within the groups should be normally distributed.

So the dependent variable could be, for example, salary, blood pressure, and study duration. These are all metric variables. And the independent variable should be nominally or ordinally scaled. For example, gender, highest level of education, or a type of therapy. Note, however, that rank order is not used with ordinal variables, so this information is lost.

13.11.6 Calculation of a two-way ANOVA

To calculate a two-way ANOVA, the following formulas are needed. Let's look at this with an example.

	Sum of squares	Degrees of freedom	Varainz	F
Total	$SS_{tot} = \sum\sum\sum (x_{mij} - \dot{x})^2$	$df_{tot} = n \cdot p \cdot q - 1$	$\sigma^2_{tot} = \frac{SS_{tot}}{df_{tot}}$	
Between	$SS_{btw} = n \cdot \sum\sum (AB_{ij} - \dot{x})^2$	$df_{btw} = p \cdot q - 1$	$\sigma^2_{btw} = \frac{SS_{btw}}{df_{btw}}$	
Factor A	$SS_A = n \cdot q \sum (A_i - \dot{x})^2$	$df_A = p - 1$	$\sigma^2_A = \frac{SS_A}{df_A}$	$F_A = \frac{\sigma^2_A}{\sigma^2_{err}}$
Factor B	$SS_B = n \cdot p \sum (B_i - \dot{x})^2$	$df_B = q - 1$	$\sigma^2_B = \frac{SS_B}{df_B}$	$F_B = \frac{\sigma^2_B}{\sigma^2_{err}}$
Interaction	$SS_{AB} = SS_{btw} - SS_A - SS_B$	$df_{AB} = (p-1) \cdot (q-1)$	$\sigma^2_{AB} = \frac{SS_{AB}}{df_{AB}}$	$F_{AB} = \frac{\sigma^2_{AB}}{\sigma^2_{err}}$
Error	$SS_{err} = \sum\sum\sum (x_{mij} - AB_{ij})^2$	$df_{err} = (n-1) \cdot p \cdot q$	$\sigma^2_{err} = \frac{SS_{err}}{df_{err}}$	

$\dot{x}$	Total mean value
A	Mean values in the groups (factor level) of factor A
B	Mean values in the groups (factor level) of factor B
AB	Mean value of the combinations of factor A and B
x_{mij}	Value m in group ij
n	Number of values per group
i	Index for groups of factor A
j	Index for groups of factor B
p	Number of groups in factor A
q	Number of groups in factor B

Let's say you work in the marketing department of a bank and you want to find out if gender and the fact that a person has studied or not have an influence on their attitude towards retirement planning.

In this example, your two independent variables (factors) are gender (male or female) and study (yes or no). Your dependent variable is attitude toward retirement planning, where 1 means "not important" and 10 means "very important."

Factor A: Study status

Factor B: Gender

	Not Studied	Studied	
Male	6	4	
	4	5	
	7	6	
	9	7	
	3	5	
Mean	5.8	5.4	5.6
Female	8	3	
	3	5	
	5	9	
	8	2	
	6	3	
Mean	6	4.4	5.2
	5.9	4.9	**5.4**

After all, is attitude toward retirement planning really a metric variable? Let's just assume that attitude toward retirement planning was measured using a Likert scale and thus we consider the resulting variable to be metric.

Mean values

In the first step we calculate the mean values of the individual groups, i.e. of male and not studied, which is 5.8 then of male and studied, which is 5.4, we now do the same for female.

Then we calculate the mean of all male and female and of not studied and studied respectively. Finally, we calculate the overall mean as 5.4.

Sums of squares

With this, we can now calculate the required sums of squares. SStot is the sum of squares of each individual value minus the overall mean.

$$SS_{tot} = \sum\sum\sum (x_{mij} - \bar{x})^2 = (6-5.4)^2 + (4-5.4)^2 + \cdots + (3-5.4)^2 = 84.8$$

Values; Total mean

$$SS_{btw} = n \cdot \sum\sum (AB_{ij} - \bar{x})^2 = 5 \cdot \left((5.8-5.4)^2 + (5.4-5.4)^2 + \cdots + (4.4-5.4)^2\right) = 7.6$$

Group means; Total mean

$$SS_A = n \cdot q \sum (A_i - \bar{x})^2 = 5 \cdot 2\left((5.9-5.4)^2 + (4.9-5.4)^2\right) = 5$$

Group mean values Factor A; Total mean

$$SS_B = n \cdot p \sum (B_j - \bar{x})^2 = 5 \cdot 2\left((5.6-5.4)^2 + (5.2-5.4)^2\right) = 0.8$$

Group mean values Factor B; Total mean

$$SS_{AB} = SS_{btw} - SS_A - SS_B = 7.6 - 5 - 0.8 = 1.8$$

$$SS_{err} = \sum\sum\sum (x_{mij} - AB_{ij})^2 = (6-5{,}8)^2 + (4-5{,}4)^2 + \cdots + (3-4{,}4)^2 = 77{,}2$$

Values; Group mean

SSbtw results from the sum of squares of the group means minus the overall mean multiplied by the number of values in the groups.

The sums of squares of the factors SSA and SSB result from the sum of squares of the means of the factor levels minus the total mean.

Now we can calculate the sum of squares for the interaction. These are obtained by calculating SSbtw minus SSA minus SSB.

Finally, we calculate the sum of squares for the error. This will calculate similar to the total sum of squares, so again we use each individual value. Only in this case, instead of subtracting the overall mean from each value, we subtract the respective group mean from each value.

Degrees of freedom

The required degrees of freedom are as follows:

$$df_{\text{tot}} = n \cdot p \cdot q - 1 = 5 \cdot 2 \cdot 2 - 1 = 19$$

$$df_{\text{btw}} = p \cdot \text{q} - 1 = 2 \cdot 2 - 1 = 3$$

$$df_{\text{A}} = \text{p} - 1 = 2 - 1 = 1$$

$$df_{\text{B}} = \text{q} - 1 = 2 - 1 = 1$$

$$df_{AB} = (\text{p} - 1) \cdot (\text{q} - 1) = 1 \cdot 1 = 1$$

$$df_{\text{err}} = (n - 1) \cdot p \cdot q = 4 \cdot 2 \cdot 2 = 16$$

Mean squares or variance

Together with the sums of squares and the degrees of freedom, the variance can now be calculated:

$$\sigma^2_{tot} = \frac{QS_{tot}}{df_{tot}} = \frac{84.8}{19} = 4.46$$

$$\sigma^2_{btw} = \frac{QS_{btw}}{df_{btw}} = \frac{7.6}{3} = 2.53$$

$$\sigma^2_{A} = \frac{QS_{A}}{df_{A}} = \frac{5}{1} = 5$$

$$\sigma^2_{B} = \frac{QS_{B}}{df_{B}} = \frac{0.8}{1} = 0.8$$

$$\sigma^2_{AB} = \frac{QS_{AB}}{df_{AB}} = \frac{1.8}{1} = 1.8$$

$$\sigma^2_{err} = \frac{QS_{err}}{df_{err}} = \frac{77.2}{16} = 4.83$$

F value

And now we can calculate the F values. These are obtained by dividing the variance of factor A, factor B or the interaction AB by the error variance.

$$F_A = \frac{\sigma_A^2}{\sigma_{err}^2} = \frac{5}{4.83} = 1.04$$

$$F_B = \frac{\sigma_B^2}{\sigma_{err}^2} = \frac{0.8}{4.83} = 0.17$$

$$F_{AB} = \frac{\sigma_{AB}^2}{\sigma_{err}^2} = \frac{1.8}{4.83} = 0.373$$

p-value

To calculate the p-value, we need the F-value, the degrees of freedom and the F-distribution. We use the F-distribution p-value calculator on DATAtab. Of course, you can also just calculate the example completely with DATAtab, more about that in the next section.

$F_A = 1.04$ $df_{\text{A}} = 1$ $df_{\text{err}} = 16$

F critical	alpha	df numerator	df denominator
4.494 =	0.05	1	16

p-Value	F-Value	df numerator	df denominator
0.323 =	1.04	1	16

$F_B = 0.17$ $df_{\text{B}} = 1$ $df_{\text{err}} = 16$

F critical	alpha	df numerator	df denominator
4.494 =	0.05	1	16

p-Value	F-Value	df numerator	df denominator
0.686 =	0.17	1	16

$F_{AB} = 0.373$ $df_{\text{AB}} = 1$ $df_{\text{err}} = 16$

F critical	alpha	df numerator	df denominator
4.494 =	0.05	1	16

p-Value	F-Value	df numerator	df denominator
0.55 =	0.373	1	16

This gives us a p-value of 0.323 for Factor A, a p-value of 0.686 for Factor B, and a p-value of 0.55 for the interaction. None of these p-values is less than 0.05 and thus we retain the respective null hypotheses.

13.11.7 Calculating two-way ANOVA with DATAtab

Calculate the example directly with DATAtab for free:

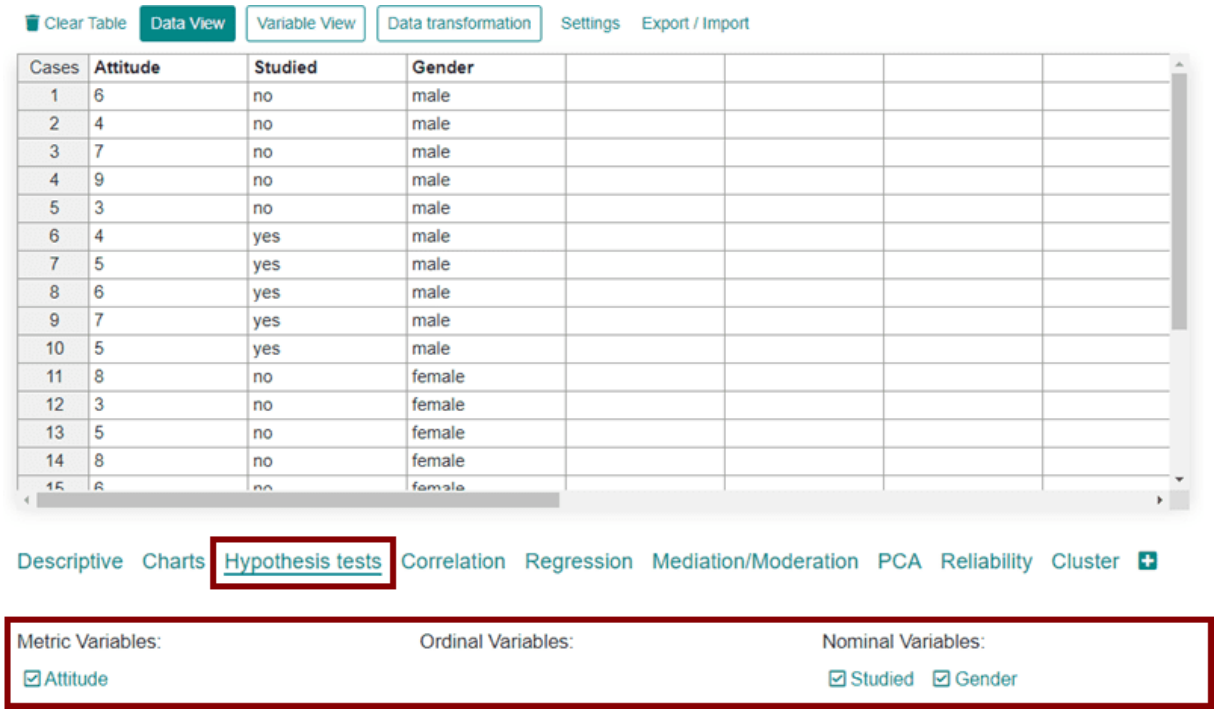
We take the same example from above. The data is now arranged in the form so that your statistics software can do something with it. In each row is a respondent.

Attitude towards retirement planning	Studied	Gender
6	no	male
4	no	male
5	no	female
...	...	...
5	yes	female
9	yes	female
2	yes	female
3	yes	female

This example consists of only 20 cases, which of course is not much, giving us very low test power, but as an example it should fit.

That's how it works with DATAtab*:*

- To calculate a two factorial analysis of variance online
- Simply visit datatab.com and copy your own data into this table.
- Then click on „hypothesis tests".
- Under this tab you will find a lot of hypothesis tests and depending on which variable you click on, you will get an appropriate hypothesis test suggested.



When you copy your data into the table, the variables appear under the table, if the correct scale level is not automatically detected, you can simply change it under Variable View.

We want to know if gender and whether you have studied or not has an impact on your attitude towards retirement planning. So we just click on all three variables.

DATAtab will now automatically calculate a two-way analysis of variance without repeated measures. DATAtab outputs the three null and the three alternative hypotheses, then the descriptive statistics and the Levene test of equality of variance. With the Levene test you can check if the variances within the groups are equal. The p-value is greater than 0.05, so we assume equality of variance within groups for these data.

Hypothesen

Copy Word Copy Excel

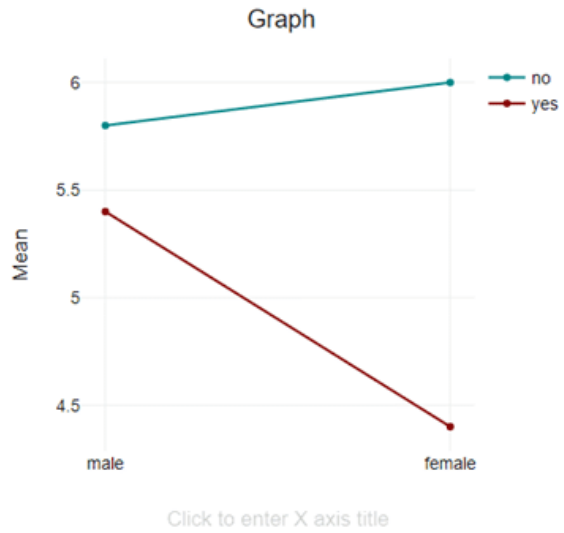
Null hypotheses	Alternative hypotheses
There is no significant difference between the groups of the independent variable Studied in relation to the dependent variable Attitude.	There is a significant difference between the groups of the independent variable Studied in relation to the dependent variable Attitude.
There is no significant difference between the groups of the independent variable Gender in relation to the dependent variable Attitude.	There is a significant difference between the groups of the independent variable Gender in relation to the dependent variable Attitude.
There is no significant interaction between the two variables Studied and Gender in relation to the dependent variable Attitude.	There is a significant interaction between the two variables Studied and Gender in relation to the dependent variable Attitude.

Descriptive statistics

Copy Word Copy Excel

Studied	Gender	Mean	Std. Deviation	N
no	male	5.8	2.39	5
	female	6	2.12	5
	Total	5.9	2.13	10
yes	male	5.4	1.14	5
	female	4.4	2.79	5
	Total	4.9	2.08	10

Download png Download svg Settings



Levene test of variance equality

Copy Word Copy Excel

F	df1	df2	p
0.98	3	16	.427

Next come the results of the two way ANOVA.

	Type III Sum of Squares	df	Mean Squares	F	p
Corrected Model	7.6	3	2.53	0.53	.671
Intercept	583.2	1	583.2	120.87	<.001
Studied	5	1	5	1.04	.324
Gender	0.8	1	0.8	0.17	.689
Studied x Gender	1.8	1	1.8	0.37	.55
Error	77.2	16	4.83		
Total	668	20			
Corrected total variation	84.8	19			

13.11.8 Interpreting two-way ANOVA

The most important in this table are the three marked rows. With these three rows, you can test whether the 3 null hypotheses we made earlier are kept or rejected.

The first row tests you null hypothesis of whether studied or not studied has an effect on attitude towards retirement planning. The second row tests whether gender has an effect on attitude. Finally the third row tests, the interaction between studied and gender.

You can read the p-value in each case right at the last column. Let's say we set the significance level at 5%. If our calculated p-value is less than 0.05, then the null hypothesis is rejected, and if the calculated p-value is greater than 0.05, the null hypothesis is not rejected.

Thus, in this case, we see that all three p-values are greater than 0.05 and thus we cannot reject any of the three null hypotheses.

Therefore, neither whether one has studied or not nor gender has a significant effect on attitudes toward retirement planning. And there is also no significant interaction between studied and gender in terms of attitudes toward retirement planning.

If you don't know exactly how to interpret the results, you can also just click on Summary in Words. In addition, it is important to check in advance whether the assumptions for the analysis of variance are met at all.

13.11.9 Interaction effect

But what exactly does interaction mean? Let us first have a look at this diagram.

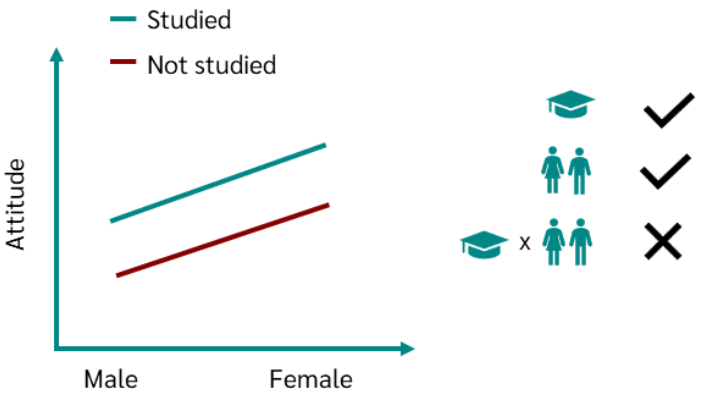


Figure 91: Interaction effect part 1

The dependent variable is plotted on the y axis, in our example the attitude towards retirement provision. On the x axis, one of the two factors is plotted, let's just take gender. The other factor is represented by lines with different colors. Green is studied and red is not studied.

The endpoints of the lines are the mean values of the groups, e.g. male and not studied.

In this diagram, one can see that both gender and the variable of having studied or not have an influence on attitudes toward retirement planning. Females have a higher value than males and studied have a higher value than not studied.

But now finally to the interaction effects, for that we compare these two graphs.

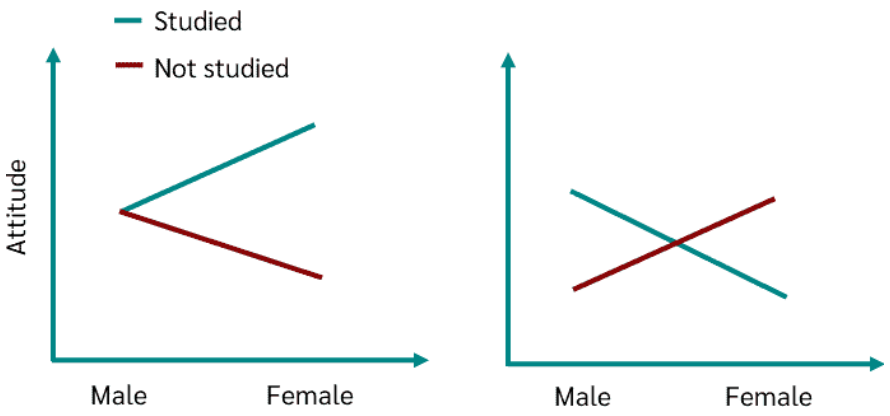


Figure 92: Interaction effect part 2

In the first case, we said there is no interaction effect. If a person has studied, he has a value that is, say, 1.5 higher than a person who has not studied.

This increase of 1.5 is independent of whether the person is male or female.

It is different in this case, here studied persons also have a higher value, but how much higher the value is depends on whether one is male or female. If I am male, there is a difference of, let's say for example 0.5 and if I am female, there is a difference of 3.5.

So in this case we clearly have an interaction between gender and study because the two variables affect each other. It makes a difference how strong the influence from studying is depending on whether I am male or female.

In this case, we do have an interaction effect, but the direction still remains the same. So females have higher scores than males and studied have higher scores than non-studied.

13.12 Two-way ANOVA with measurement repetition

If we look at the most common types of the analysis of variance, we distinguish once between the one-way and the two-way analysis of variance and on the other side the analysis of variance without measurement repetition and with measurement repetition. Now we will look at the two-way analysis of variance with measurement repetition.

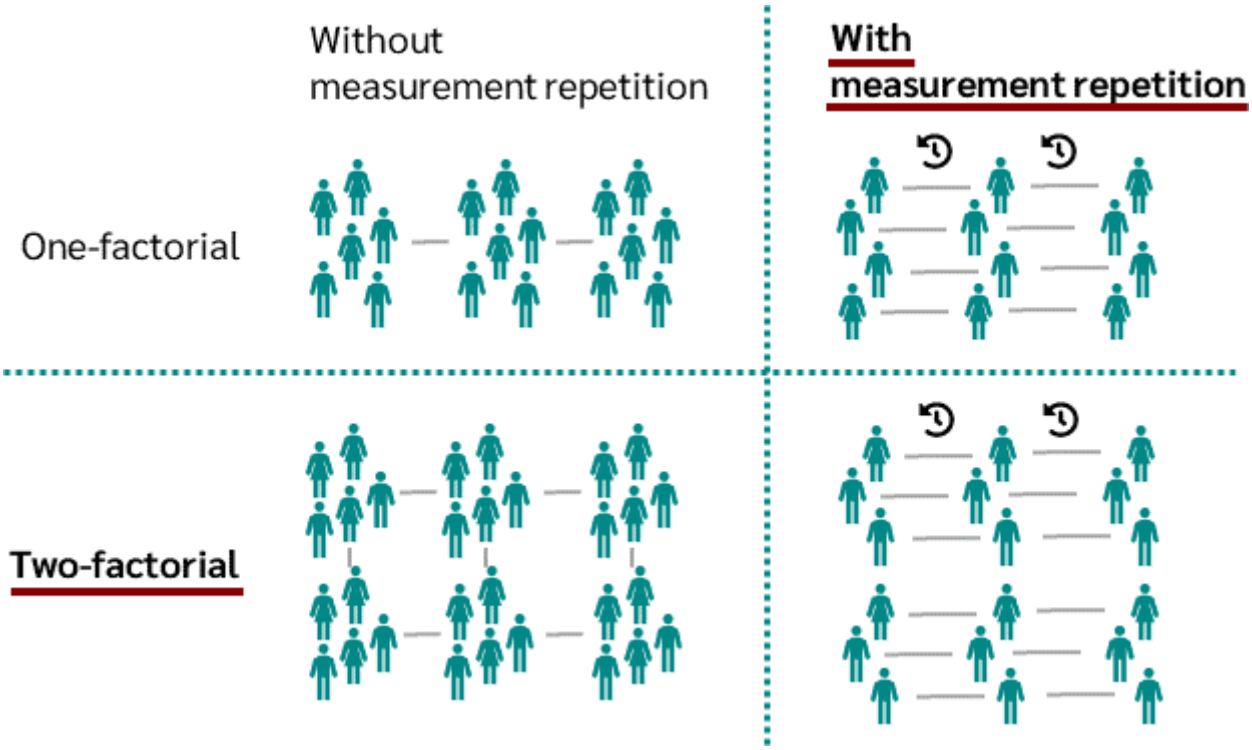
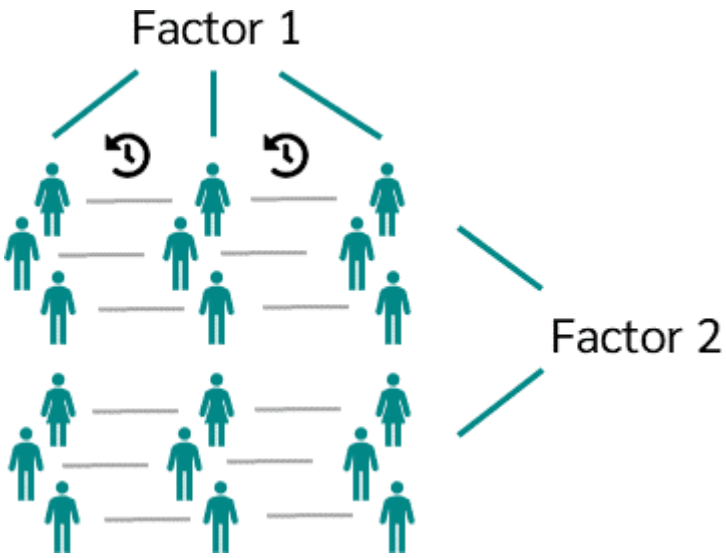


Figure 93: Two-way ANOVA with measurement repetition

Two-way analysis of variance with measurement repetition tests whether there is a difference between more than two samples divided between two variables or factors.

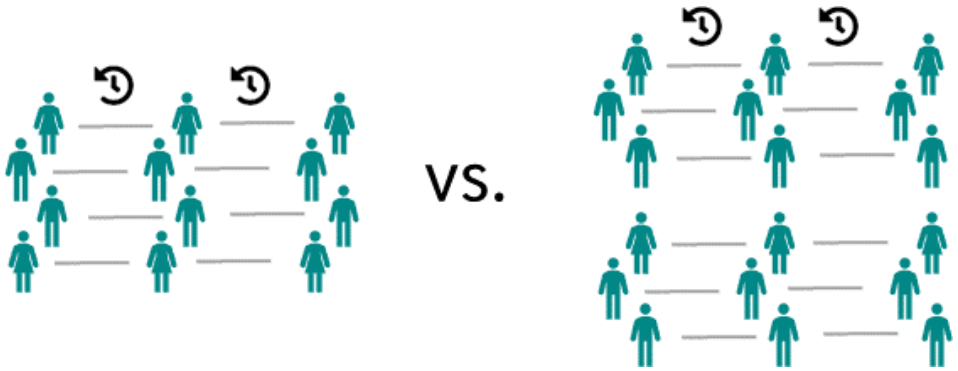
In contrast to the two-factorial analysis of variance without measurement repetitions, one of the factors is thereby created by measurement repetitions. In other words, one factor is a dependent sample.



13.12.1 Sample with measurement repetition

What is the difference to the "normal" one-factor analysis of variance with repeated measures? Or what is the difference between one-factorial and two-factorial?

Single factorial ANOVA with repeated measures tests whether there are statistically significant differences between three or more dependent samples.



In a dependent sample, the measured values are linked. Thus, one and the same person is measured at several time points.

13.12.2 Example two-way ANOVA with repeated measures

For example, if you take a sample of people with high blood pressure and measure their blood pressure before, during and after treatment, this is a dependent sample. This is because the same person is interviewed at different times.

You may want to know if the treatment for high blood pressure has an effect on the blood pressure. So you want to know if blood pressure changes over time.

But what if you have different therapies and you want to see if there is a difference between them? You now have two factors, one for the therapy and one for the repeated measurements. Since you now have two factors and one of the factors is a dependent sample, you use a two-way repeated measures analysis of variance.

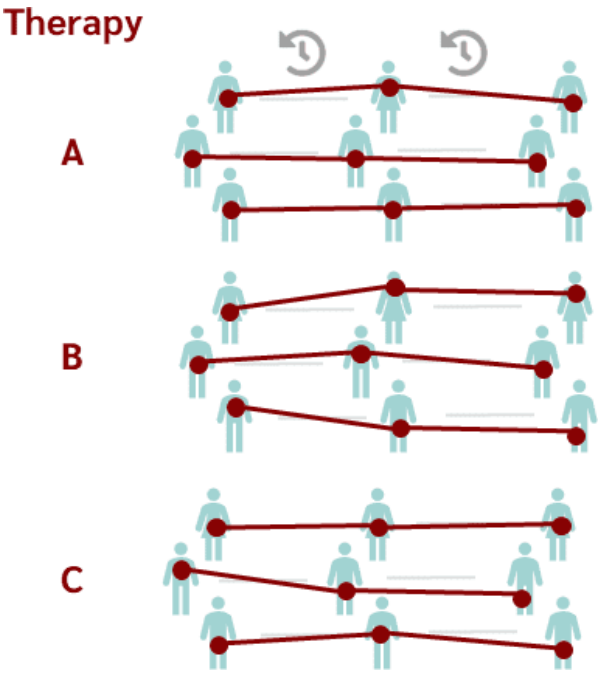


Figure 94: Example two-way ANOVA with measurement repetition

Using two-way analysis of variance with repeated measures, you can now answer three things:

- Does the first factor with measurement repetition have an effect on the dependent variable?
- Does the second factor have an effect on the dependent variable?
- Is there an interaction between factor 1 and factor 2?

13.12.3 Hypotheses

As already indicated, you can test three statements with the 2 factorial analysis of variance, so there are also 3 null hypotheses and therefore also 3 alternative hypotheses.

Null hypotheses:

- The mean values of the different measurement times do not differ (There are no significant differences between the "groups" of the first factor).
- The mean values of the different groups of the second factor do not differ.
- One factor has no influence on the effect of the other factor

13.12.4 Assumptions of the two-way analysis of variance with repeated measures

In order for a two-way analysis of variance with measurement repetition to be calculated, the following prerequisites must be met:

- The scale level of the dependent variable should be metric. For example, salary or blood pressure.
- The scale level of the factors should be categorical.
- The measurements of one factor should be dependent, e.g. the measurements should have arisen from repeated measurements of the same person.
- The measurements from the other factor should be independent, i.e. the measurement from one group should not be influenced by the measurement from another group.

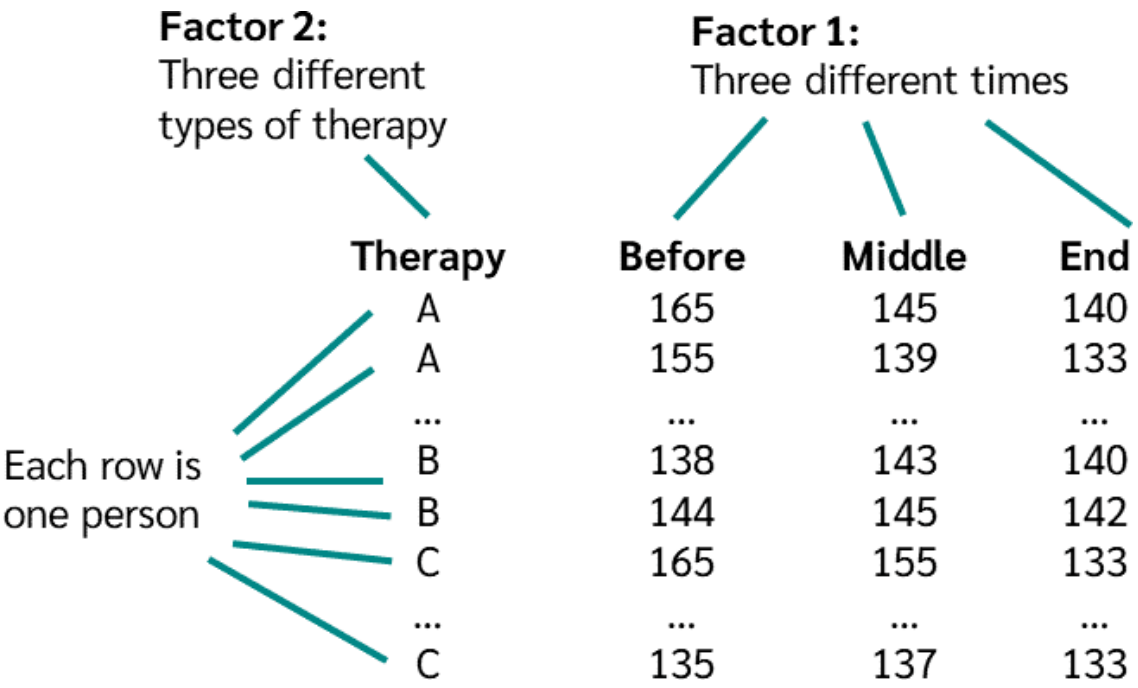
- The variances in each group should be approximately equal. This can be checked with the Levene test
- The data within the groups should be normally distributed.
- The variances of the differences between all combinations of the different groups should be equal (Sphericity). This assumption can be tested using Mauchly's test of sphericity.

13.12.5 Calculate two-way ANOVA with repeated measures

That's how it works with DATAtab*:*

If you want to calculate the example directly with DATAtab, just download the example data from datatab.net.

Let's say this is our data we want to analyze. Each row is one person, the first factor reflects the three time points before therapy in the middle and at the end of therapy, and the second factor reflects the type of therapy.



To calculate a two-way analysis of variance with repeated measures online, simply visit datatab.net and copy your own data into the table.

Then click on hypothesis testing. Under this tab you will find a lot of hypothesis tests and depending on which variable you click on, you will get an appropriate hypothesis test suggested.

When you copy your data into the table, a list of the inserted variables appear below it. If the correct scale level is not automatically detected, you can easily change it under *Variables View*.

For example, if we click on "Before", "Middle" and "End", an analysis of variance with repeated measures is automatically calculated. But we also want to include the therapy, so we just click on "Therapy".

Now we get a two-way analysis of variance with measurement repetition.

We can read the three null and the three alternative hypotheses. Then we get the descriptive statistics output and then the results of the analysis of variance are displayed. We will look at these in detail in a moment.

13.12.6 Interpret two-way analysis of variance with repeated measures

Most important in this table are the plotted three rows, with these three rows, you can test if the 3 null hypotheses we made before are kept or rejected. The first row tests you null hypothesis, whether blood pressure changes over time, so whether the therapies have an effect on blood pressure.

ANOVA

Copy Word Copy Excel

	Sum of squares	df	Mean Squares	F	p
Before, Middle, End	524.93	2	262.47	5.18	.014
Therapy	49.6	2	24.8	0.12	.889
A x B	134.67	4	33.67	0.66	.623
Between	2,575.33	14	183.95		
Within the sample	2,525.73	12	210.48		
Residuum	1,217.07	24	50.71		
Within	1,876.67	30	62.56		
Total	4,452	44	101.18		

The second row tests whether there is a difference between the respective therapies with respect to blood pressure. And the last row checks if there is an interaction between the two factors.

You can read the p-value at the very back of each one. Let's say we set the significance level at 5%. If our calculated p-value is less than 0.05, then the respective null hypothesis is rejected and if the calculated p-value is greater than 0.05, then the null hypothesis is not rejected.

Thus, we see that the p-value of before, middle and end is less than 0.05 and thus the before, middle and end times are significantly different in terms of blood pressure. The p-value in the second row is greater than 0.05, so the therapies have the same mean values over time.

It is important to note that the mean value over the three time points is considered here. It could also be that in one therapy the blood pressure increases and in the other therapy the blood pressure decreases, but on average over the time points the blood pressure is the same, then we would not get a significant difference here.

If that were the case, however, we would have an interaction between the therapies and time. We test this with the last hypothesis.

In this case, there is no significant interaction between therapy and time.

If you don't know exactly how to interpret the results, you can also just click on Summary in Words.

13.13 Kruskal-Wallis test

The Kruskal-Wallis test (H-test) is a hypothesis test for multiple independent samples that is used when the requirements for a single factor analysis of variance are not met.

Since the Kruskal-Wallis test is a nonparametric test (also called a distribution-free procedure), the data used do not have to be normally distributed, in contrast to the analysis of variance. The only requirement is that the data have ordinal scale level.

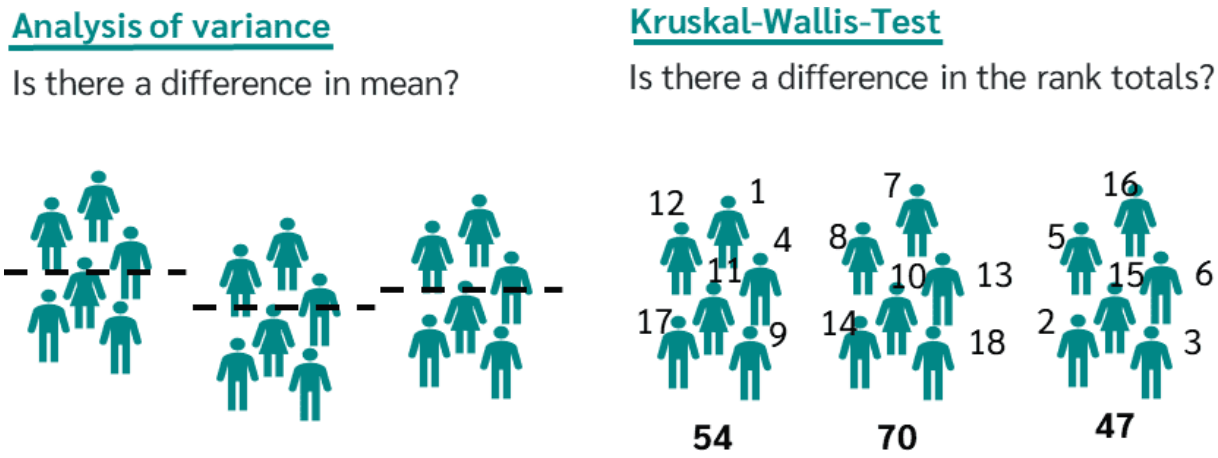


Figure 95: Analysis of variance and Kruskal-Wallis test

In the Kruskal-Wallis test, ordinal variables are sufficient, since non-parametric procedures do not use the differences of the values, but the ranks (which value is larger, which is smaller). Therefore, the Kruskal-Wallis test is also often called rank variance analysis according to Kruskal and Wallis.

13.13.1 Examples for the Kruskal-Wallis test

For the Kruskal-Wallis test, of course, the same examples can be used as for the single factor analysis of variance, but with the addition that the data need not be normally distributed.

- Medical example:

For a pharmaceutical company you want to test whether a drug XY has an influence on body weight. For this purpose, the drug is administered to 20

test persons, 20 test persons receive a placebo and 20 test persons receive no medication or placebo.

- Social science example:

Do three age groups differ in terms of daily television viewing?

13.13.2 Research question and hypotheses in the Kruskal-Wallis test

The research question in the Kruskal-Wallis test may be: Is there a difference in the central tendency of several independent samples? This question then gives rise to the null and alternative hypotheses.

Null hypothesis

The independent samples all have the same central tendency and therefore come from the same population.

Alternative hypothesis

At least one of the independent samples does not have the same central tendency as the others and therefore comes from a different population.

13.13.3 Assumptions of the Kruskal-Wallis test

To compute a Kruskal-Wallis test, only several independent random samples with at least ordinally scaled characteristics must be available. The variables do not have to satisfy a distribution curve.

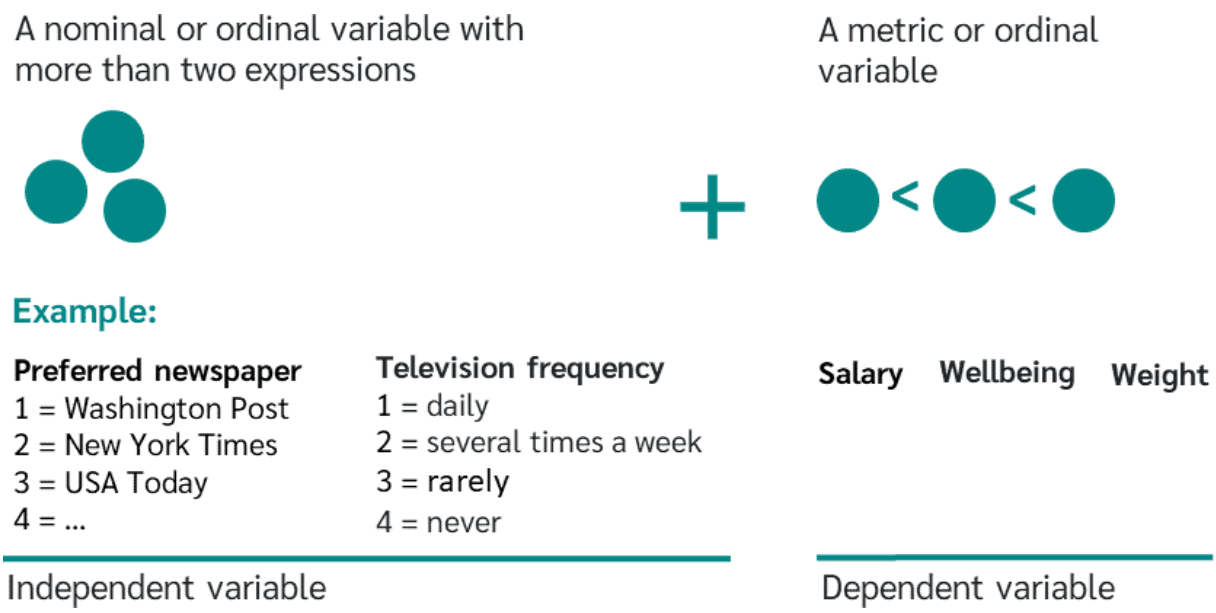


Figure 96: Assumptions of the Kurskal-Wallis test

13.13.4 Calculate Kruskal-Wallis test

The calculation of the Kruskal and Wallis rank variance analysis is similar to that of the Mann-Whitney U test, which is the nonparametric counterpart of the t test for independent samples.

Group	Response time	Rank
A	34	2
A	36	4
A	41	7
A	43	9
B	44	10
B	37	5
B	45	11
B	33	1
C	35	3
C	39	6
C	42	8
C	46	12

Rank sums:

$R_A = 2 + 4 + 7 + 9 = 22$

$R_B = 10 + 5 + 11 + 1 = 27$

$R_C = 3 + 6 + 8 + 12 = 29$

Mean Rank Sum:

$\bar{R}_A = 22 / 4 = 5.5$

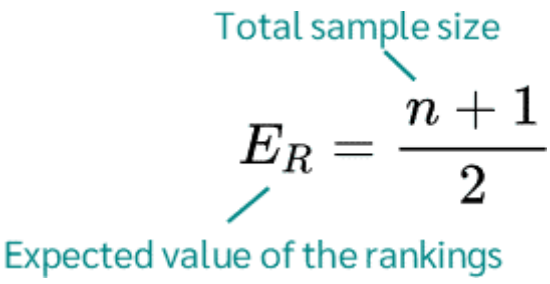
$\bar{R}_B = 27 / 4 = 6.75$

$\bar{R}_C = 29 / 4 = 7.25$

$$E_R = \frac{n+1}{2} = \frac{12+1}{2} = 6.5$$

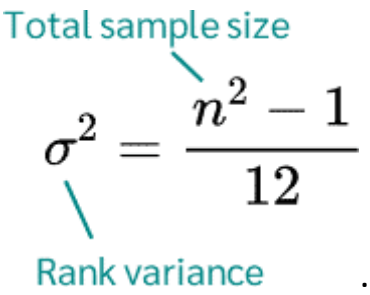
Let's say the null hypothesis holds and thus there is no difference between the independent samples. Then high and low ranks are distributed randomly across the samples and should be equally distributed across the groups. Therefore, the probability that a rank is assigned to a group is the same for all groups (Bühner & Ziegler ,2017).

If there is no difference between the groups, the mean value of the rankings should also be the same in all groups. The expected value of the rankings for each group is then given by



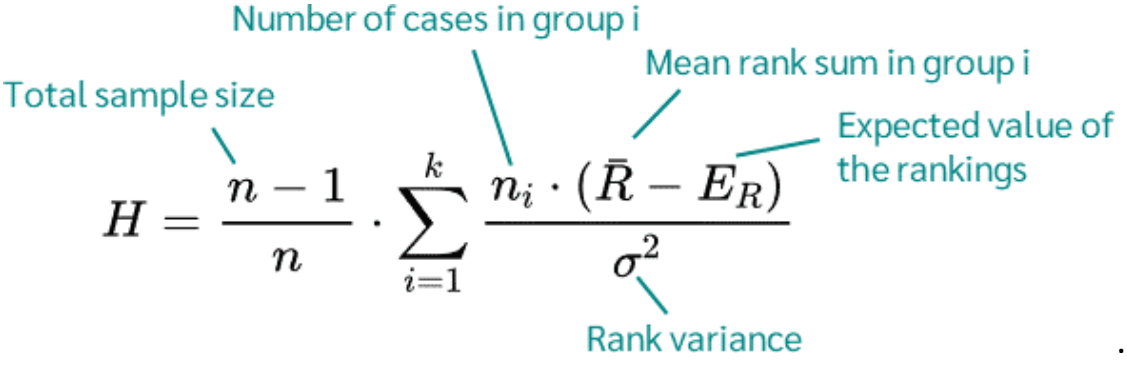
.

Each sample thus has the same expected value of the ranks, which corresponds to the expected value of the population. Furthermore, the variance of the ranks is needed, this can be calculated with the following formula:



.

In the Kruskal-Wallis test, the test variable H is calculated. The H value corresponds to the χ2 value. The H value results from:


Number of cases in group i
Mean rank sum in group i
Total sample size
Expected value of the rankings
$H = \frac{n-1}{n} \cdot \sum_{i=1}^{k} \frac{n_i \cdot (\bar{R} - E_R)}{\sigma^2}$
Rank variance

.

The critical H value can be read from the table of critical χ2 values.

13.13.5 Kruskal-Wallis test example

You have measured the reaction time of three groups and want to know if there is a difference between them.

First, we assign a rank to each person, then we calculate the rank sum and the mean rank sum.

Group	Response time	Rank
A	34	2
A	36	4
A	41	7
A	43	9
B	44	10
B	37	5
B	45	11
B	33	1
C	35	3
C	39	6
C	42	8
C	46	12

Rank sums:

$R_A = 2 + 4 + 7 + 9 = 22$

$R_B = 10 + 5 + 11 + 1 = 27$

$R_C = 3 + 6 + 8 + 12 = 29$

Mean Rank Sum:

$\overline{R}_A = 22 / 4 = 5.5$

$\overline{R}_B = 27 / 4 = 6.75$

$\overline{R}_C = 29 / 4 = 7.25$

$$E_R = \frac{n+1}{2} = \frac{12+1}{2} = 6.5$$

We measured reaction time in twelve people, so the number of cases is twelve. The degrees of freedom are given by the number of groups minus one, so we have two degrees of freedom.

Number of cases

$n = 12$

Degrees of freedom

$df = 2$

Rank variance

$$\sigma_R^2 = \frac{n^2 - 1}{12} = \frac{12^2 - 1}{12} = 11.92$$

Expected value of the rankings

$E_R = 6.5$

Mean Rank Totals:

$\overline{R}_A = 22 / 4 = 5.5$

$\overline{R}_B = 27 / 4 = 6.75$

$\overline{R}_C = 29 / 4 = 7.25$

Now we have calculated all values to calculate the test quantity H.

Test statistic H — equivalent to χ^2

$$H = \frac{n - 1}{12} \cdot \sum_{i=1}^{k} \frac{n_i (\bar{R}_i - E_R)^2}{\sigma_R^2}$$

$$H = \frac{12 - 1}{12} \cdot 4 \frac{(5.5 - 6.5)^2 + (6.75 - 6.5)^2 + (7.25 - 6.5)^2}{11.92}$$

$$= 0.5$$

After the H-value or the chi-square value has been calculated, the critical chi-square value can be read from the table of critical chi-square values.

Table of chi-squared distribution

Significance level Alpha	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01
Degrees of freedom								
1	0	0.001	1.642	2.706	3.841	5.024	5.412	6.635
2	0.01	0.051	3.219	4.605	5.991	7.378	7.824	9.21
3	0.072	0.216	4.642	6.251	7.815	9.348	9.837	11.345
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086

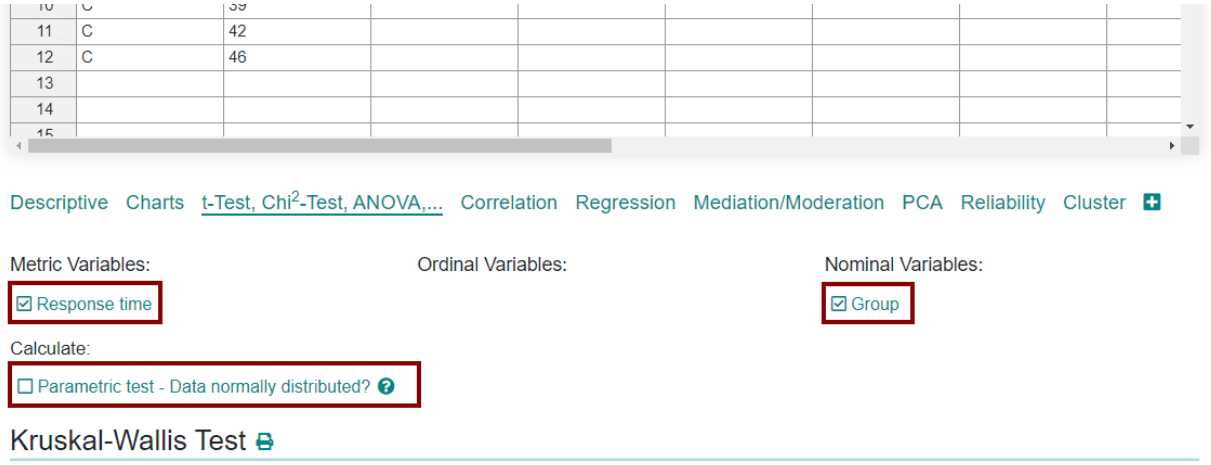
Therefore, at a significance level of 5%, the critical chi-square value is 5.991. This critical value is therefore larger than the calculated chi-square or H-value. Thus, the null hypothesis is maintained and there is no difference in reaction time in the three groups.

That’s how it works with DATAtab*:*

Of course you can calculate the Kruskal-Wallis test online with DATAtab.

- Just go to the statistics calculator,
- copy your data into the table and
- select the tab "Hypothesis tests”.

Then you just must select the variables you want to analyze and uncheck "Parametric test".



14. Statistical methods for testing correlations

The following chapter deals with the statistical testing of correlations and will address correlation analyses as well as partial correlations.

14.1 Correlation

Correlation analysis is a statistical technique that gives you information about the relationship between **metric or ordinal variables.**

Correlation analysis can be calculated to investigate the relationship of variables. How strong the correlation is is determined by the correlation coefficient, which varies **from -1 to +1.**

Correlation analyses can thus be used to make a statement about the strength and direction of the correlation.

For example, one question might be: "Is there a relationship between the age at which a child speaks his or her first sentences and later school success?"

14.1.1 Correlation and causality

If the correlation analysis shows that two characteristics are related, it can subsequently be tested whether one characteristic can be used to predict the other characteristic. Thus, if a relationship given in the example is detected it can be tested whether school success can be predicted by the age at which a child speaks his or her first sentences by means of a linear regression.

But beware! Correlations need not be causal relationships. Any correlations that are discovered should therefore be investigated more closely, but never interpreted immediately in terms of content, even if this would be obvious.

14.1.2 Correlation and causality example

If the correlation between sales figures and price is analyzed and a strong correlation occurs, it would be quite logical to assume that sales figures are influenced by price (and not vice versa), but this assumption can by no means be proven based on a correlation analysis.

Furthermore, it can happen that the correlation between variable x and y is generated by **variable z**, see Partial Correlation for more information.

14.1.3 Correlation interpretation

With the help of correlation analysis two statements can be made:

- one about the direction
- and one about the strength

of the linear relationship between two metric or ordinally scaled variables. The direction indicates whether the correlation is positive or negative, while the strength indicates whether the correlation between the variables is strong or weak.

14.1.4 Direction of correlation

Positive correlation

A positive correlation exists when larger values of variable A are accompanied by larger values of variable B. Height and shoe size, for example, correlate positively, resulting in a correlation coefficient that lies between 0 and 1, i.e., assumes a positive value.

Negative correlation

A negative correlation exists when larger values of variable A are accompanied by smaller values of variable B. The price of a product and its

sales volume usually have a negative correlation. This means that the more expensive a product is, the lower its sales volume. In this case, the correlation coefficient is between -1 and 0, i.e., it takes on a negative value.

14.1.5 Strength of correlation

Regarding the strength of the correlation, the following table can be used as a guide:

Table 2: Strength of the correlation

Amount of r	**Strength of the correlation**
0,0 < 0,1	no correlation
0,1 < 0,3	low correlation
0,3 < 0,5	medium correlation
0,5 < 0,7	high correlation
0,7 < 1	very high correlation

Source: Kuckartz et al., 2013, p. 213

14.1.6 Scatter plot and correlation

Just as important as the consideration of the correlation coefficient is the graphical consideration of the correlation of two variables in a scatter diagram.

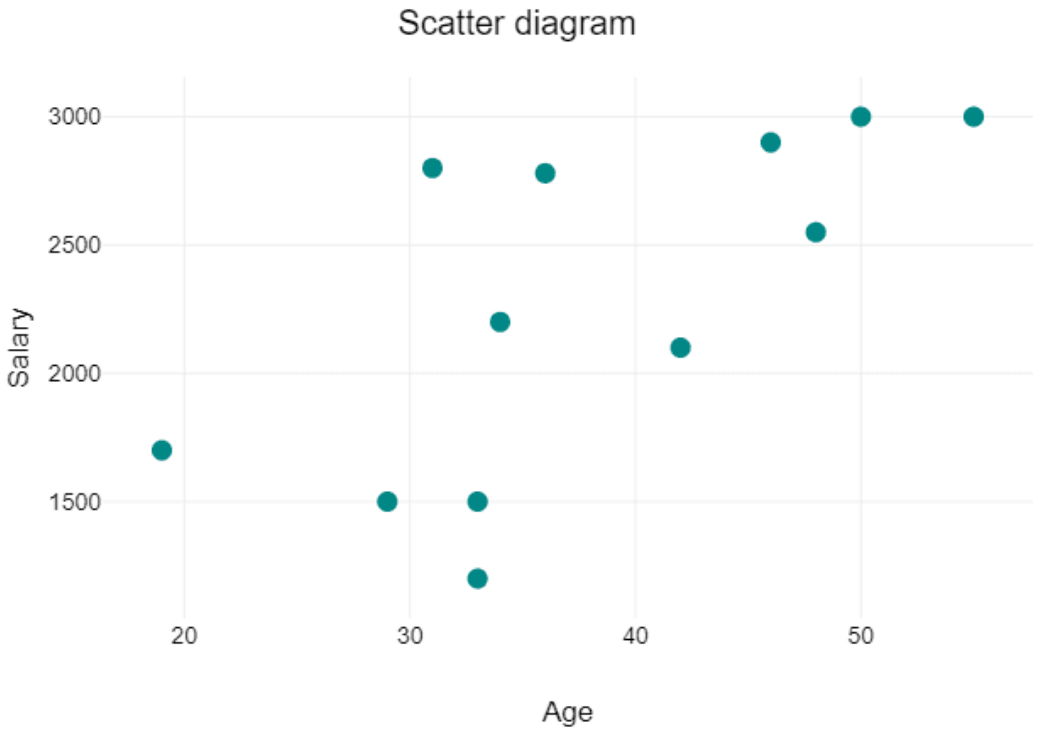


Figure 97: Scatter plot and correlation

The scatter plot gives you a rough estimate of whether there is a correlation, whether it is linear or nonlinear, and whether there are outliers.

14.1.7 Test correlation for significance

If there is a correlation in the sample, it is still necessary to test whether there is enough evidence that the correlation also exists in the population. Thus, the question arises when a correlation coefficient can be considered statistically significant.

The significance of correlation coefficients can be checked using a **t-test. As** a rule, this involves analyzing whether the correlation coefficient differs significantly from zero. Linear independence is thus tested. In this case, the null hypothesis is that there is no correlation in the population. In contrast, the alternative hypothesis assumes that there is a correlation.

As with any other hypothesis test, the significance level is first set, usually at 5%. If the calculated p-value is **now** below 5%, the null hypothesis is rejected and the alternative hypothesis applies. In this case, it is therefore assumed that there is a correlation between the variables in the population.

The t-value for testing the hypothesis is given by

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

,

where *n is* the sample size and *r is* the determined correlation in the sample. The corresponding p-value can be easily calculated in the correlation calculator on DATAtab.

14.1.8 Directional and non-directional hypotheses

With correlation analysis you can test directional and non-directional correlation hypotheses.

- **Non-directional correlation hypothesis:**

You are only interested in whether there is a relationship or correlation between two variables, for example, whether there is a correlation between age and salary, but you are not interested in the direction of this correlation.

- **Directional correlation hypothesis:**

You are also interested in the direction of the correlation, i.e. whether there is a positive or negative correlation between the variables.

Your alternative hypothesis is then e.g. age has a positive influence on salary. What you have to pay attention to in the case of a directional hypothesis, we will go through at the bottom of the example.

14.2 Pearson correlation analysis

With the Pearson correlation analysis you get a statement about the linear correlation between metric scaled variables. The respective covariance is used for the calculation.

The covariance gives a positive value if there is a positive correlation between the variables and a negative value if there is a negative correlation. The covariance is calculated as:

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

.

However, the covariance is not standardized and can assume values between plus and minus infinity. This makes it difficult to compare the strength of relationships between different variables. For this reason, the correlation coefficient, also called **product-moment correlation coefficient**, is calculated.

The correlation coefficient is obtained by normalizing the covariance. For this normalization, the variances of the two variables involved are used and the correlation coefficient is calculated as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

.

The **Pearson correlation coefficient** can take values between -1 and +1 and can be interpreted as follows:

- The **value +1** means that there is an entirely positive linear relationship (the more, the more).
- The **value -1** indicates that there is an entirely negative linear relationship (the more, the less).
- At a **value of 0, there** is no linear relationship, the variables do not correlate with each other.

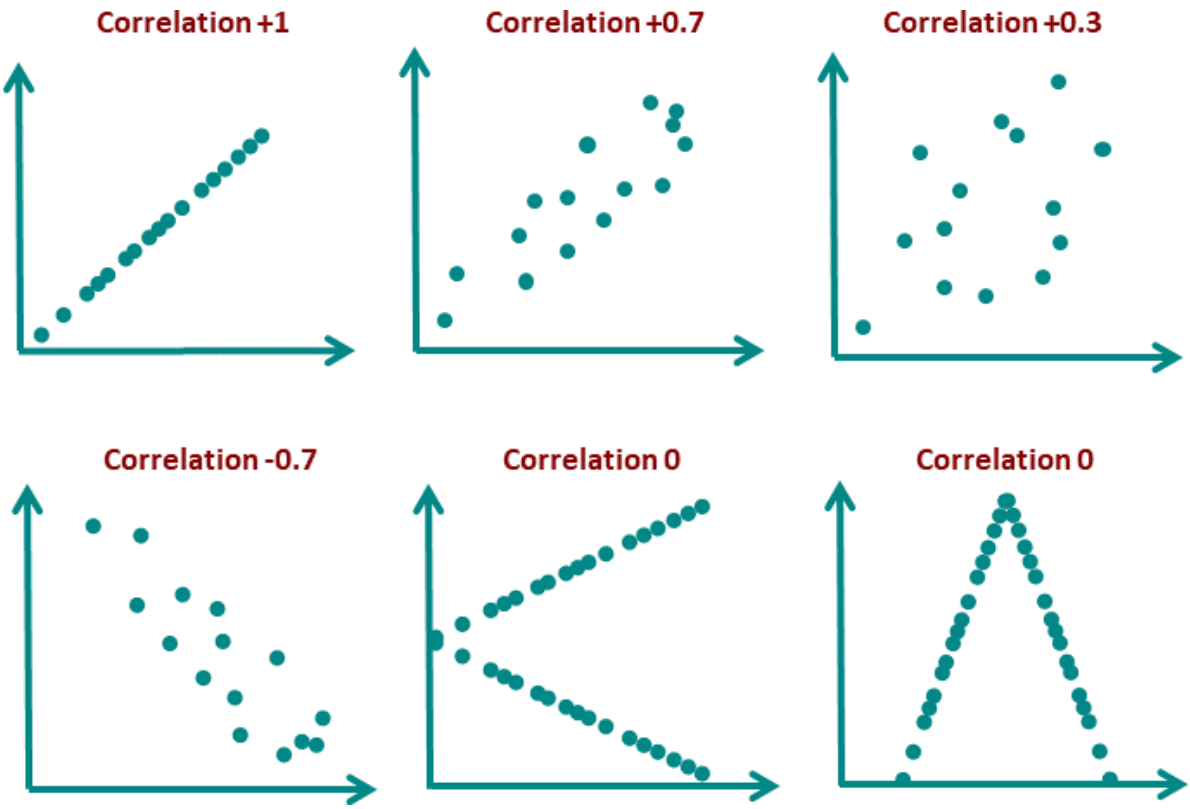


Figure 98: Strength and direction of the correlation coefficients

Now finally the strength of the relationship can be interpreted. This can be illustrated by the following table:

$\|r\|$	Strength of correlation
0.0 < 0,1	no correlation
0.1 < 0,3	little correlation
0.3 < 0,5	medium correlation
0.5 < 0,7	high correlation
0.7 < 1	very high correlation

To check in advance whether there is a linear relationship, **scatter diagrams** should be viewed. In this way, the respective correlation between the variables can also be checked visually in advance. Interpreting the Pearson correlation is only meaningful and purposeful in the presence of **linear relationships.**

Finally, the calculated correlation coefficient can also be tested for **significance.** This serves to find out whether the correlation found also applies to the population. If an **undirected** hypothesis is to be tested, the null hypothesis could be that there is no correlation between two characteristics in the population. In the **directed** case, it can be tested whether there is either a positive or a negative correlation in the population.

Example: There is a positive correlation in the population and therefore the following alternative hypothesis is formulated: "The greater a person's climate awareness is, the greater is his or her sustainability awareness."

To calculate the probability that a correlation detected in the sample also exists in the population, a test variable is required. In mathematical terms, this test variable follows the t-distribution with n-2 degrees of freedom (df).

Finally, the **test variable** can be used to decide whether the null hypothesis is retained or rejected, i.e. whether or not there is also a positive correlation between climate awareness and sustainability awareness in the population.

14.2.1 Pearson Correlation assumptions

For Pearson correlation to be used, the variables must be normally distributed and there must be a linear relationship between the variables. The normal distribution can be tested either analytically or graphically with the Q-Q plot. Whether the variables have a linear correlation is best checked with a scatter plot. If these conditions are not met, then the Spearman correlation is used.

If these described conditions are not met, then the **Spearman correlation is** used, which will be discussed in more detail in the next chapter.

14.3 Spearman rank correlation

Spearman correlation analysis is used to calculate the relationship between two variables that have ordinal level of measurement. Spearman rank correlation is the non-parametric equivalent of Pearson correlation analysis. This procedure is therefore used when the prerequisites for a correlation analysis (=parametric procedure) are not met, i.e. when there is no metric data and no normal distribution. In this context it is often referred to as "Spearman correlation" or "Spearman's Rho" if Spearman rank correlation is meant.

The questions that can be treated by Spearman rank correlation are similar to those of the Pearson correlation coefficient, i.e. "Is there a correlation between two variables or characteristics". For example: "Is there a correlation between age and religiousness in the France population?

The calculation of the **rank correlation** is based on the ranking system of the data series. This means that the measured values are not used for the calculation, but are transformed into ranks. The test is then performed using these ranks.

For the **rank correlation coefficient** ρ, values between -1 and 1 are possible. If there is a value less than zero ($\rho < 0$), there is a negative linear correlation. If a value greater than zero ($\rho > 0$), there is a positive linear relationship. If the value is zero ($\rho = 0$), there is no relationship between the variables. As with the Spearman correlation coefficient, the strength of the correlation can be classified as follows:

	Strength of the correlation
0,0 < 0,1	no correlation
0,1 < 0,3	little correlation
0,3 < 0,5	medium correlation
0,5 < 0,7	high correlation
0,7 < 1	very high correlation

14.4 Point biserial correlation

The point biserial correlation is used when one of the variables is dichotomous, e.g. studied and not studied, and the other has metric scale level, e.g. salary.

The calculation of a point biserial correlation is the same as the calculation of the Pearson correlation. To calculate it, one of the two expressions of the dichotomous variable is coded as 0 and the other as 1.

14.5 Partial correlation

The **partial correlation**, also called **partial correlation**, calculates the correlation between two variables excluding a third variable. Thus, it can be found out whether the correlation r_{xy} between variable x and y is generated by the variable z.

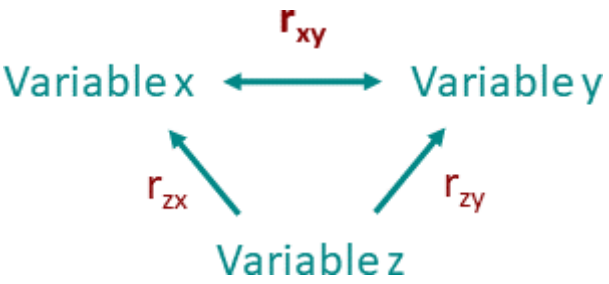


Figure 99: Partial correlation

The partial correlation $r_{xy,z}$ thus tells how strongly variable x correlates with variable y when the correlation of both variables with variable z is calculated.

14.5.1 Calculation of the partial correlation

For the calculation of the partial correlation, the three correlations between the individual variables are required. The partial correlation is then calculated as follows:

$$r_{xy,z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2) \cdot (1 - r_{yz}^2)}}$$

- r_{xy} = Correlation between variable x and y
- r_{xz} = Correlation of the third variable z with the variable x
- r_{yz} = Correlation of the third variable z with the variable y

r_{xy}	r_{xz}	r_{yz}	$r_{xy,z}$
0.63	0.57	0.88	**= 0.329**

14.5.2 Partial correlation example

Probably the most prominent example of partial correlation is that of storks and babies by Robert Matthews "Storks Deliver Babies", although the following figures are fictitious. The correlation between the number of **nesting storks** and the **birth rate** is r=0.63. This result surprised scientists at first, because they had not expected any correlation.

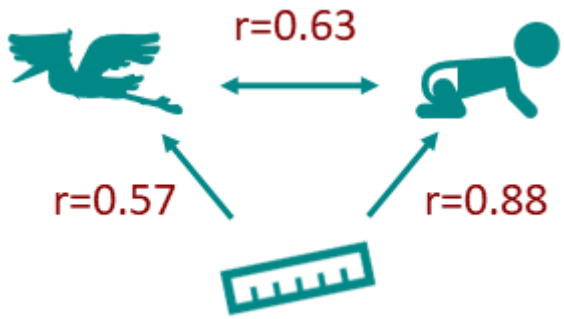


Figure 100: Bogus correlation storks and birth rate

What has been calculated in this example is a typical **spurious correlation** obtained when the correlation of two variables is generated by a **third** variable. Both the number of nesting storks (r=0.57) and the birth rate (r=0.88) correlate with the area per inhabitant. To control for area per inhabitant, only samples with equal area per inhabitant could be taken. However, this is often not feasible because the sample size becomes too small.

Another way is to calculate a correlation unaffected by the area per inhabitant. To obtain this **partial correlation**, the individual correlations are substituted into the equation:

$$0,329 = \frac{0,63 - 0,57 \cdot 0,88}{\sqrt{(1 - 0{,}57^2) \cdot (1 - 0{,}88^2)}}$$

.

The result shows that the number of nesting storks and the birth rate are only correlated by 0.329 when the area is excluded.

14.5.3 Partial correlation 2nd order

A 2nd order partial correlation occurs when two variables are removed from the correlation of two **variables rather** than just one. The equation to calculate the 2nd order partial correlation is:

$$r_{xy,z1z2} = \frac{r_{xy,z1} - r_{xz2,z1} \cdot r_{yz2,z1}}{\sqrt{(1 - r^2_{xz2,z1}) \cdot (1 - r^2_{yz2,z1})}}$$

where x and y are the two variables of which one wants to know the correlation excluding variables z1 and z2.

14.5.4 Example: Pearson correlation

A student wants to know if there is a **correlation** between the height, weight, and age of the participants in the statistics course. For this purpose, the student drew a **sample**, which is described in the table below.

Height	**Weight**	**Age**
1,62	53	20
1,72	71	30
1,85	85	25
1,82	86	24
1,72	76	23
1,55	62	25
1,65	68	26

Height	Weight	Age
1,77	77	20
1,83	97	33
1,53	65	24

This is how it works with DATAtab*:*

To analyze the linear relationships of these characteristics by means of a correlation analysis, you can calculate a correlation with DATAtab. First copy the table above into the statistics calculator.

Then click on "Correlation" and select the three variables from the example. Finally, you get the following table and scatter plot to visually illustrate the correlations:

Pearson Correlation Analysis

Test assumptions | Effect size | Summary in words

Hypotheses

Copy Word | Copy Excel

Null hypothesis	Alternative hypothesis
There is no association between Body height and Weight	There is a association between Body height and Weight

Valid cases

Copy Word | Copy Excel

	Valid cases
Number	10

Correlation

Copy Word | Copy Excel

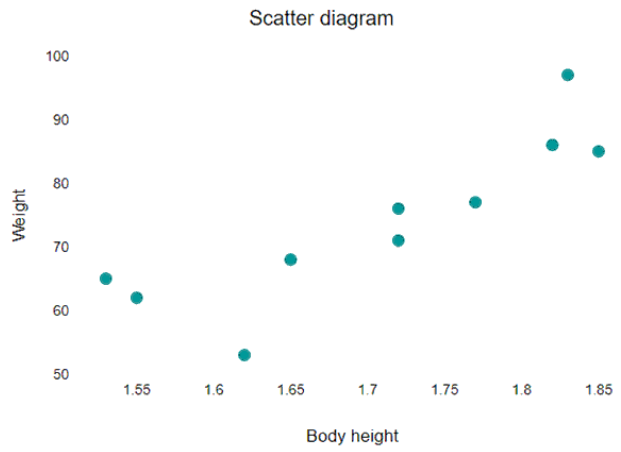
	r	p (2-tailed)
Body height and Weight	0.86	.001

Scatter plot

Size of the graphic

- small
- medium
- large

Download png | Download svg | Settings



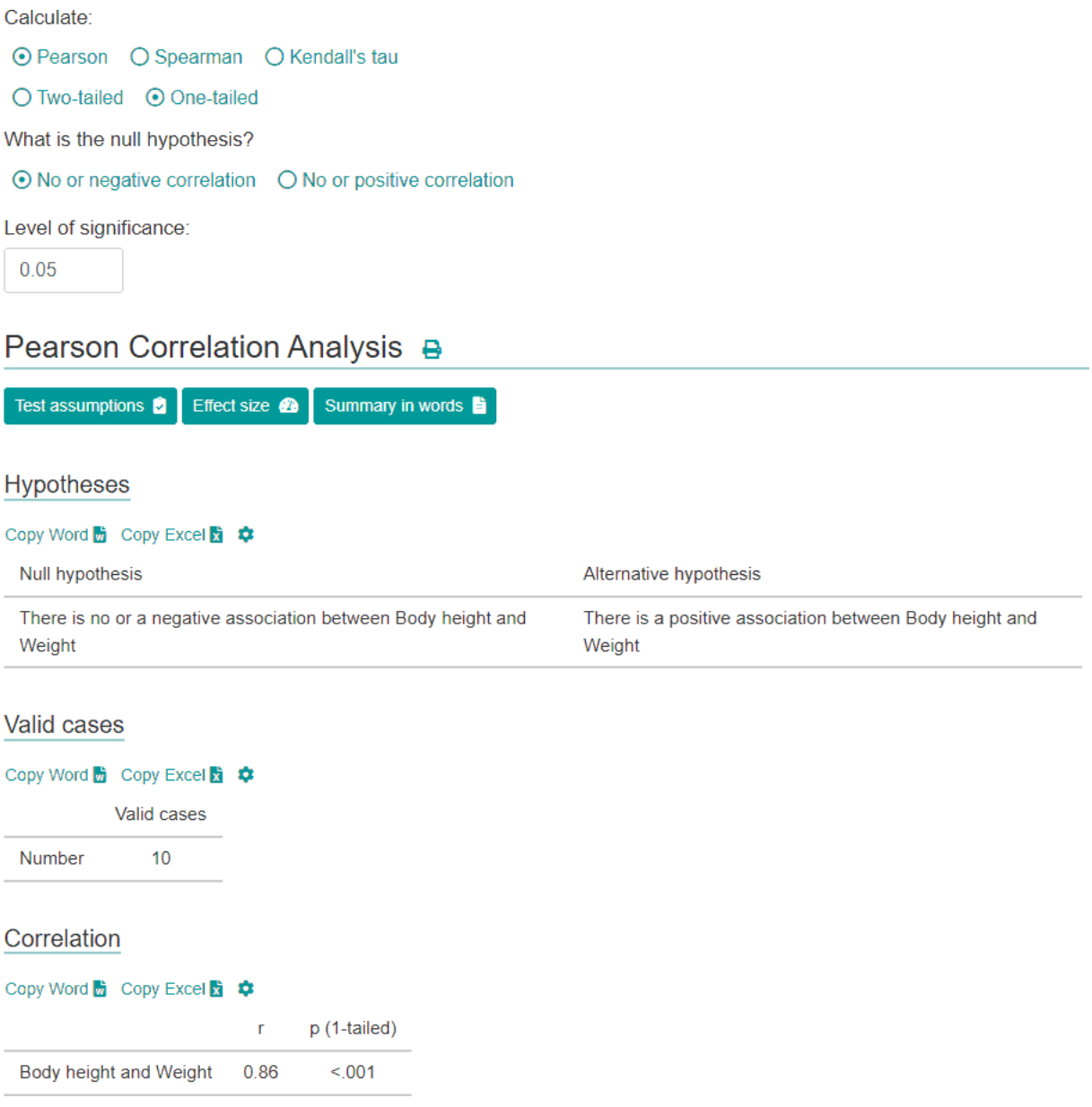
First, you will get the null and the alternative hypothesis. The null hypothesis is: "There is no correlation between height and weight". Then you get the correlation coefficient and the p value. If you click on Summary in words, you will get the following interpretation:

A Pearson correlation analysis was performed to test whether there is a relationship between height and weight. The result of the Pearson correlation analysis showed that there was a significant relationship between height and weight, r(8) = 0.86, p = 0.001.

There is a very high, positive correlation between the variables of height and weight, r= 0.86. Thus, there is a very high, positive correlation in this sample between height and weight.

14.5.5 Directional (one-sided) correlation hypothesis

Of course, in DATatab you can also choose to calculate a directional hypothesis. This is how it works:



15. Regression Analysis

What is a regression analysis, what questions does it answer and what types of regressions are there? These and other questions are the focus of the following chapter.

15.1 Basics of regression

Regression is a statistical method that allows modeling relationships between a dependent variable and one or more independent variables.

A regression analysis thus serves to infer or predict a further variable on the basis of one or more variables.

For example, you might be interested in the factors that influence a person's salary. For example, you could consider the highest level of education, the weekly working hours, and the age of a person.

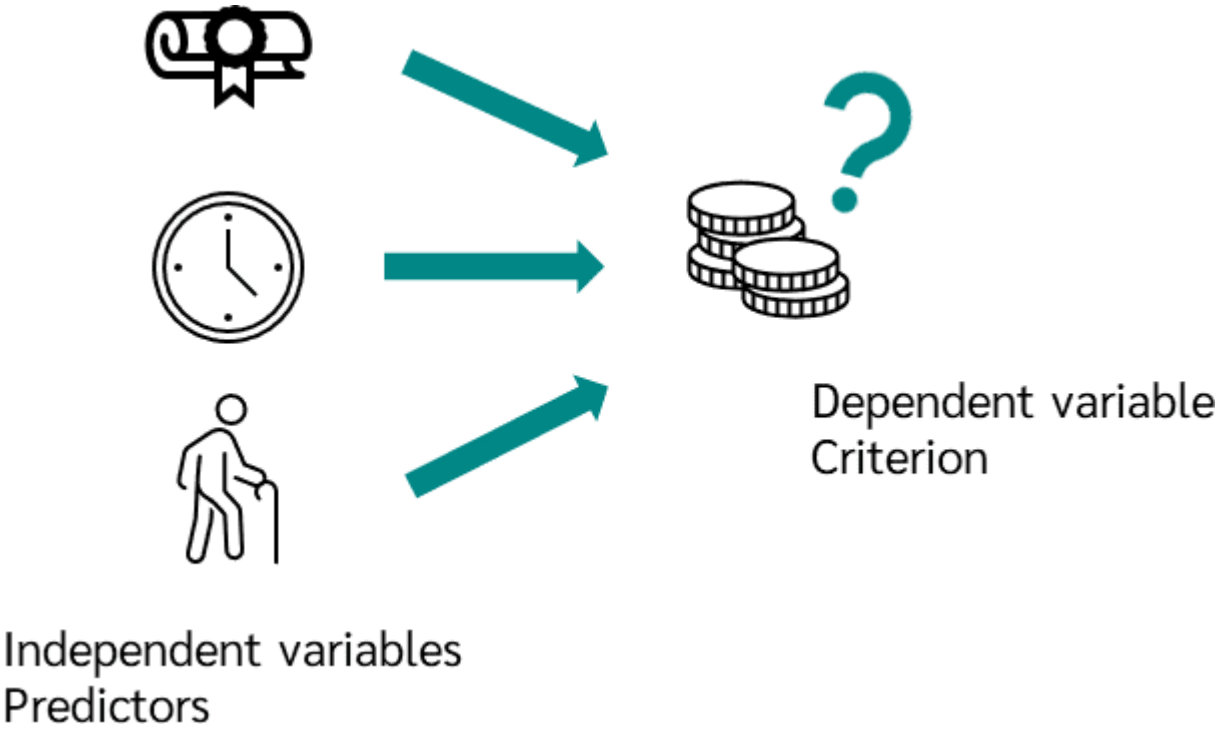


Figure 101: Question of the regression

Further you could now investigate whether these three variables have an influence on a person's salary. If so, you can predict a person's salary by using the highest education level, the weekly working hours and the age of a person.

The variable to be inferred is called the **dependent variable** (criterion). The variables that are used for prediction are called **independent variables** (predictors).

In the example above, salary is the dependent variable, and the highest level of education, weekly working hours and age are the independent variables.

15.1.1 Using a regression analysis

By performing a regression analysis two goals can be pursued. On the one hand, the influence of one or more variables on another variable can be measured, and on the other hand, the regression can be used to predict a variable by one or more other variables. For example:

1) Measurement of the influence of one or more variables on another variable.

- What has an impact on children's ability to concentrate?
- Do parents' educational attainment and place of residence affect children's future educational attainment?

2) Prediction of a variable by one or more other variables.

- How long does a patient stay in the hospital?
- What product is a person most likely to buy from an online store?

Regression analysis provides information on how the value of the dependent variable changes when one of the independent variables is changed.

15.1.2 Types of regression analysis

Regression analyses are divided into simple linear regression, multiple linear regression and logistic regression. Which regression analysis is used is determined on the one hand by the number of independent variables and on the other hand by the scale level of the dependent variable.

	Number of independent variables	Scale of measurement dependent variable	Scale of measurement independent variable
Simple linear Regression	one	metric	metric, ordinal, nominal
Multiple lineare Regression	multiple	metric	metric, ordinal, nominal
Logistic Regression	multiple	ordinal, nominal	metric, ordinal, nominal

If you only want to use one variable for prediction, a simple regression is used. If you use more than one variable, you need to perform a multiple regression. If the dependent variable is nominally scaled, a logistic regression must be calculated.

If the dependent variable is metrically scaled, a linear regression is used. Whether a linear or a non-linear regression is used depends on the relationship itself. In order to perform a linear regression, a linear relationship between the independent variables and the dependent variable is necessary.

Independent variable of the regression

No matter which regression is calculated, the scale level of the independent variables can take any form (metric, ordinal and nominal). However, if there is an ordinal or nominal variable with more than two values, so-called dummy variables must be formed.

15.1.3 Dummy variables and Reference category

When an independent variable is categorical, it is encoded as a set of binary dummy variables before being included in the regression model.

When dummy variables are created, a variable with several categories is made into several variables with only 2 categories each.

One of the categories is set as the reference category and a new variable is created for each of the remaining categories.

Let's take an example to illustrate this. Suppose you are studying the effect of education level (a categorical variable with three levels: high school, college, and graduate) on salary. In order to include this categorical variable in a regression model, it needs to be encoded as dummy variables.

Let's say we use high school as reference category and we create two dummy variables: is_college and is_graduate. The variable is_college for example will take a value of 1 if the individual has a college degree and 0 otherwise.

15.1.4 Examples of regression:

- **Simple linear regression**

Does the weekly working time have an influence on the hourly wage of employees?

- **Multiple linear regression**

Do the weekly working hours and the age of employees have an influence on their hourly wage?

Logistic regression

Do the weekly working hours and the age of employed people have an influence on the likelihood that they are at risk of burnout?

dependent variable

independent variables

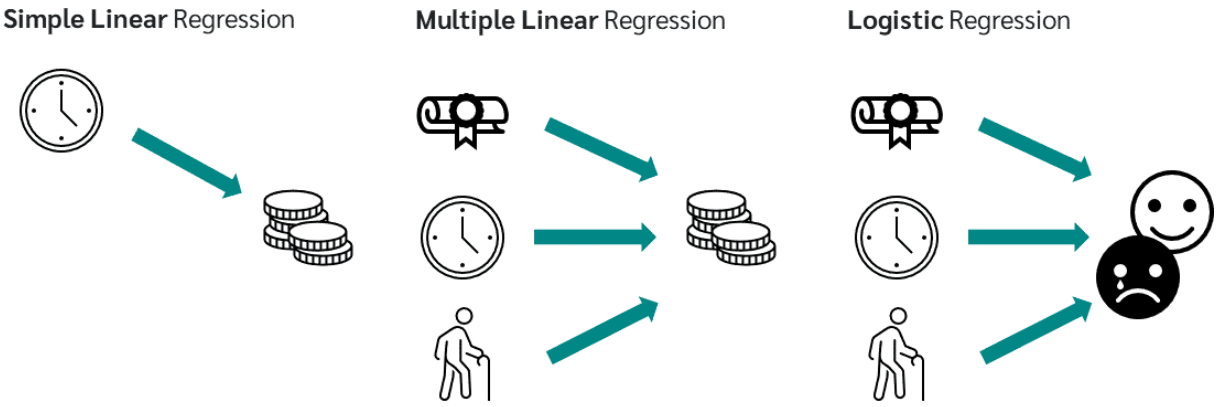


Figure 102: Types of regression

This is how it works with DATAtab:

You want to calculate a regression analysis? DATAtab requires only three simple steps:

- copy your data into the table of the statistics calculator.
- select the range of the regression.
- select a dependent variable and one or more independent variables

If one of the independent variables has a categorical scale level (ordinal or nominal), dummy variables are automatically created, and a reference category is defined. As soon as a variable contains only numbers, DATAtab's statistics calculator automatically recognizes that it is a metric variable.

15.2 Linear Regression

Linear regression analysis is used to create a model that describes the relationship between a dependent variable and one or more independent variables. Depending on whether there are one or more independent variables, a distinction is made between simple and multiple linear regression analysis.

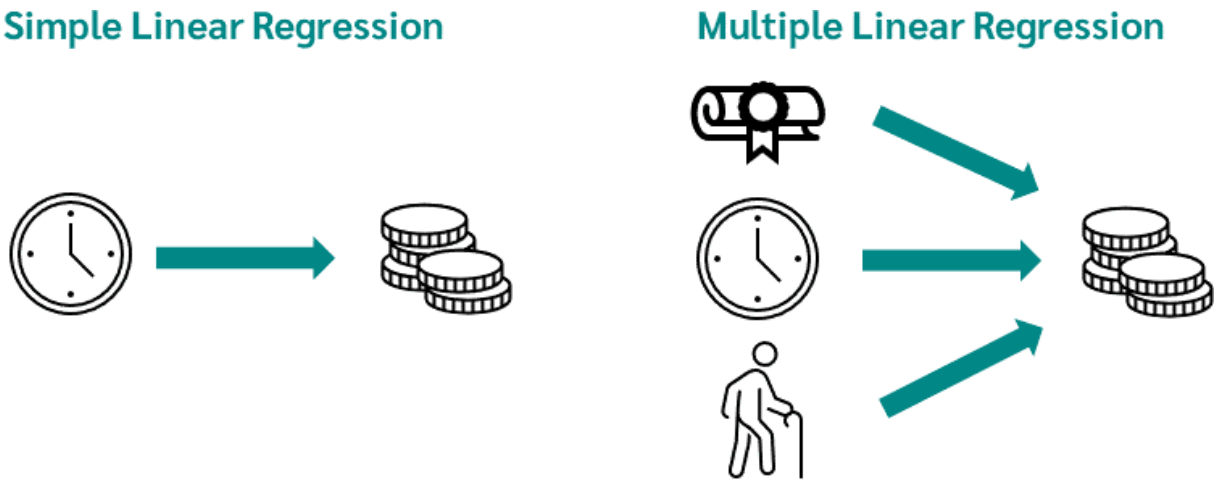


Figure 103: Simple and multiple linear regression

In linear regression, an important prerequisite is that the measurement scale of the dependent variable is metric and a normal distribution exists. If the dependent variable is categorical, a logistic regression is used. You can easily perform a regression analysis in the linear regression calculator here on DATAtab.

The prerequisite for linear regression is that the scale level of the dependent variable is interval-scaled and that there is a normal distribution. If the dependent variable is categorical, logistic regression is used.

- Example simple linear regression: Does height (IV) affect a person's weight (DV)?
- Example multiple linear regression: Do height (IV 1) and gender (IV 2) affect a person's weight (DV)?

15.2.1 Simple Linear Regression

The goal of **simple linear regression** is to predict the value of a dependent variable given an independent variable.

The greater the linear relationship between the independent variable and the dependent variable, the more accurate the prediction. This also means that the greater the proportion of the variance of the dependent variable can be explained by the independent variable.

Visually, the relationship between the variables can be shown in a scatter plot. The greater the linear relationship between the dependent and independent variables, the more the data points lie on a straight line.

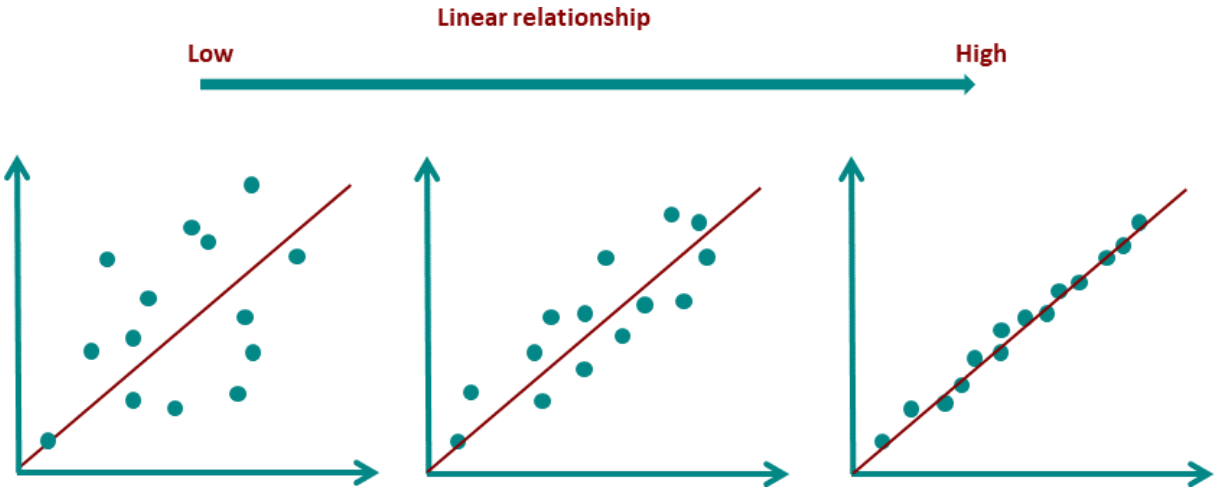
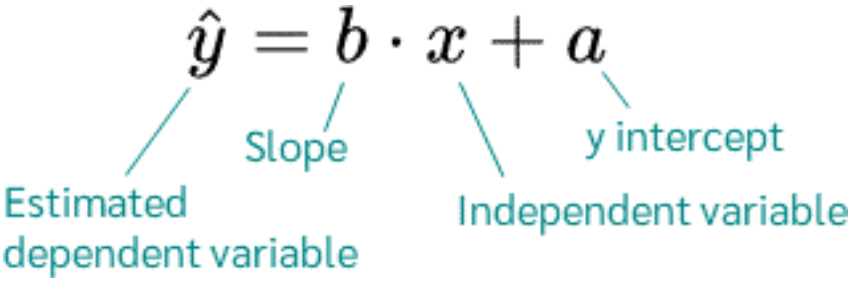


Figure 104: Scatter plot to show the correlation.

The task of simple linear regression is to exactly determine the straight line which best describes the linear relationship between the dependent and independent variable. In linear regression analysis, a straight line is drawn in the scatter plot. To determine this straight line, linear regression uses the method of least squares. The **regression line** can be described by the following **equation:**



"Regression coefficients" are understood to mean:

- **a**: The intersection point with the y-axis,
- **b**: The slope of the straight line

ŷ is the respective estimate of the y-value. This means that for each x-value the corresponding y-value is estimated. In our example, this means that the height of people is used to estimate their weight.

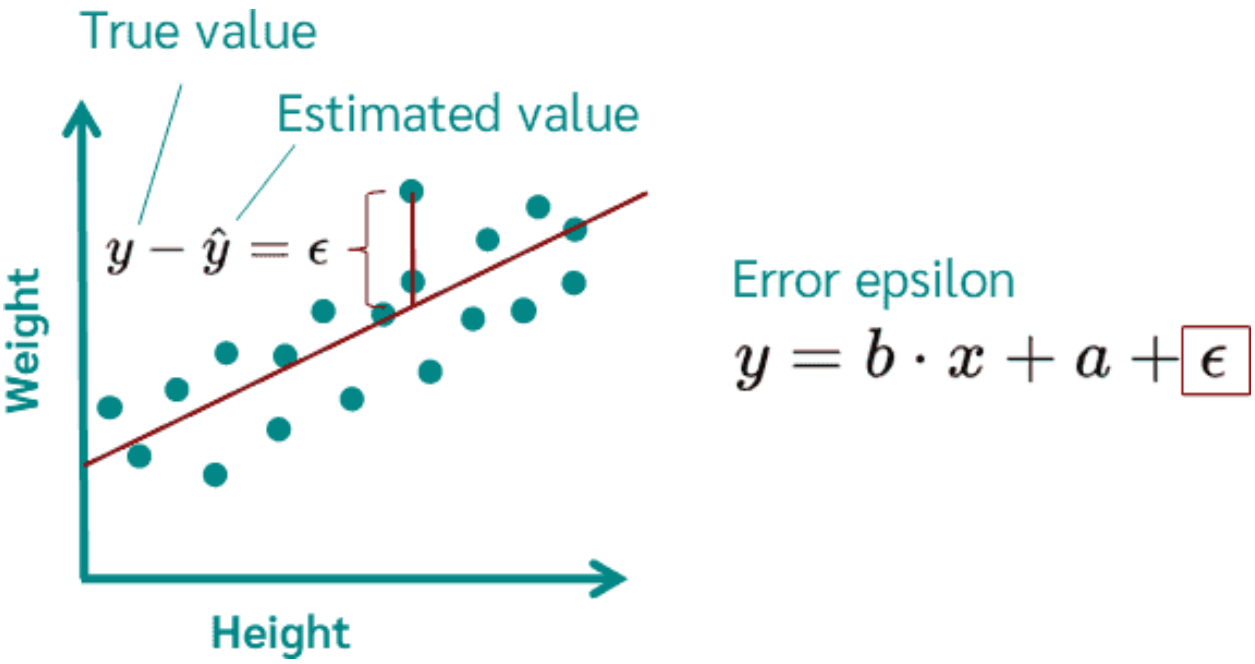


Figure 105: Representation of the regression line

If all points (measured values) were exactly on one straight line, the estimate would be perfect. However, this is almost never the case and therefore, in most cases a straight line must be found, which is as close as possible to the individual data points.

The attempt is thus made to keep the error in the estimation as small as possible so that the distance between the estimated value and the true value is as small as possible. This distance or error is called the "residual", is abbreviated as "e" (error) and can be represented by the greek letter epsilon (ϵ).

When calculating the regression line, an attempt is made to determine the regression coefficients (a and b) so that the **sum of the squared residuals** is minimal. **(OLS- "Ordinary Least Squares")**

The **regression coefficient b** can now have different signs, which can be interpreted as follows

- **b > 0**: there is a positive correlation between x and y (the greater x, the greater y)
- **b < 0**: there is a negative correlation between x and y (the greater x, the smaller y)
- **b = 0**: there is no correlation between x and y

Standardized regression coefficients are usually designated by the letter "beta". These are values that are comparable with each other. Here the unit of measurement of the variable is no longer important. The standardized regression coefficient (beta) is automatically output by DATAtab.

15.2.2 Multiple linear regression

Unlike simple linear regression, multiple linear regression allows more than two independent variables to be considered. The goal is to estimate a variable based on several other variables. The variable to be estimated is called the dependent variable (criterion). The variables that are used for the prediction are called independent variables (predictors).

Multiple linear regression is frequently used in empirical social research as well as in market research. In both areas it is of interest to find out what influence different factors have on a variable. For example, what determinants influence a person's health or purchasing behavior? Since regression analyses are used in many disciplines, you will now find some content-related examples:

An example in **marketing** might be: For a video streaming service, you want to predict how many times a month a person streams videos. For this, you get a dataset of visitor data (age, income, gender, ...).

Here is a **medical example**: You want to find out which factors have an influence on the cholesterol level of patients. To do this, you analyze a data set of patients with their cholesterol levels, age, hours of exercise per week, etc.

The equation necessary for the calculation of a multiple regression is obtained with k dependent variables as:

Simple Linear Regression

$$\hat{y} = b \cdot x + a$$

Multiple Linear Regression

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

The coefficients can now be interpreted in a similar way to the linear regression equation. If all independent variables are 0, the result is a. If an independent variable changes by one unit, the associated coefficient indicates

how much the dependent variable changes. An increase in the independent variable x_i thus increases or decreases the dependent variable y by b_i units.

15.2.2.1 Multiple Regression vs. Multivariate Regression

Multiple regression should not be confused with multivariate regression. In the former case, the influence of several independent variables on a dependent variable is examined. In the second case, several regression models are calculated to allow conclusions to be drawn about several dependent variables. Consequently, in a multiple regression, one dependent variable is taken into account, whereas in a multivariate regression, several dependent variables are analyzed.

15.2.2.2 Coefficient of determination

In order to find out how well the regression model can predict or explain the dependent variable, two main measures are used. This is on the one hand the coefficient of determination R2 and on the other hand the standard estimation error.

The coefficient of determination R2, also known as the variance explanation, indicates how large the portion of the variance is that can be explained by the independent variables. The more variance can be explained, the better the regression model is.

In order to calculate R2, the variance of the estimated value is related to the variance in the observed values:

$$R^2 = \frac{s^2_{\hat{y}}}{s^2_y}$$

Variance of the predicted values

Variance of the observed values

15.2.2.3 Adjusted R2

The coefficient of determination R2 is influenced by the number of independent variables used. The more independent variables are included in the regression model, the greater the variance resolution R2. To take this into account, the adjusted R2 is used.

$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

15.2.2.4 Standard estimation error

The standard estimation error is the standard deviation of the estimation error. This gives an impression of how much the prediction differs from the correct value. Graphically interpreted, the standard estimation error is the dispersion of the observed values around the regression line.

The coefficient of determination and the standard estimation error are used for simple and multiple linear regression.

15.2.2.5 Standardized and unstandardized regression coefficient

The regression coefficient is distinguished between the standardized and the unstandardized regression coefficient. The unstandardized regression coefficients are the coefficients that occur or are used in the regression equation and are abbreviated b.

The standardized regression coefficients are obtained by multiplying the regression coefficient bi by the standard deviation of the dependent variable Sxi and dividing by the standard deviation of the respective independent variable Sy.

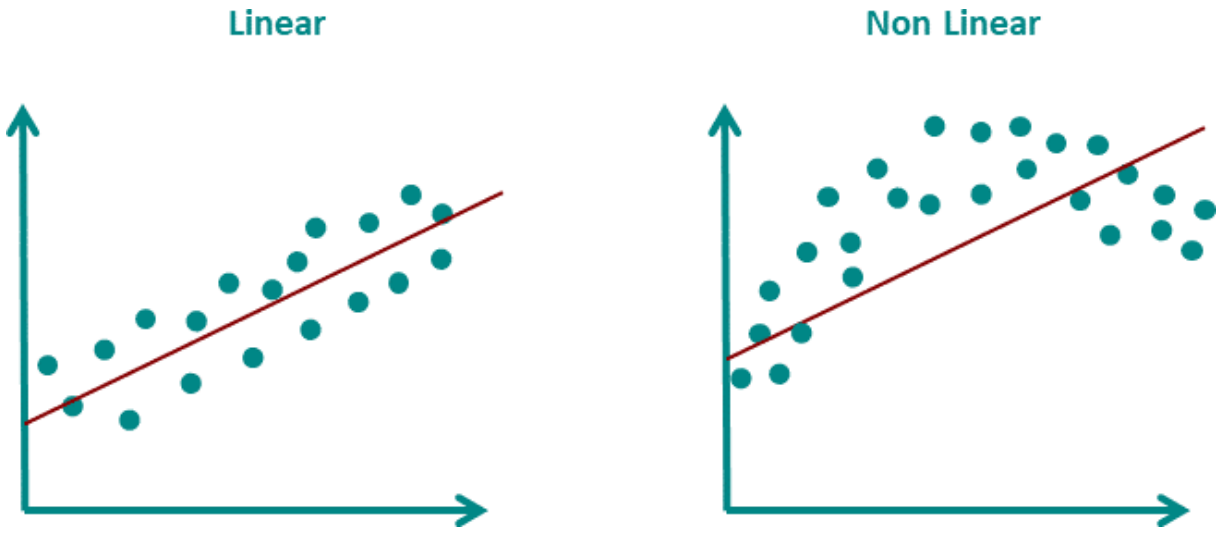
15.2.2.6 Assumptions of Linear Regression

In order to interpret the results of the regression analysis meaningfully, certain conditions must be met.

- **Linearity:** There must be a linear relationship between the dependent and independent variables.
- **Homoscedasticity:** The residuals must have a constant variance.
- **Normality:** Normally distributed error
- **No multicollinearity:** No high correlation between the independent variables
- **No auto-correlation:** The error component should have no auto correlation

15.2.2.7 Linearity

In linear regression, a straight line is drawn through the data. This straight line should represent all points as good as possible. If the points are distributed in a non-linear way, the straight line cannot fulfill this task.

- 

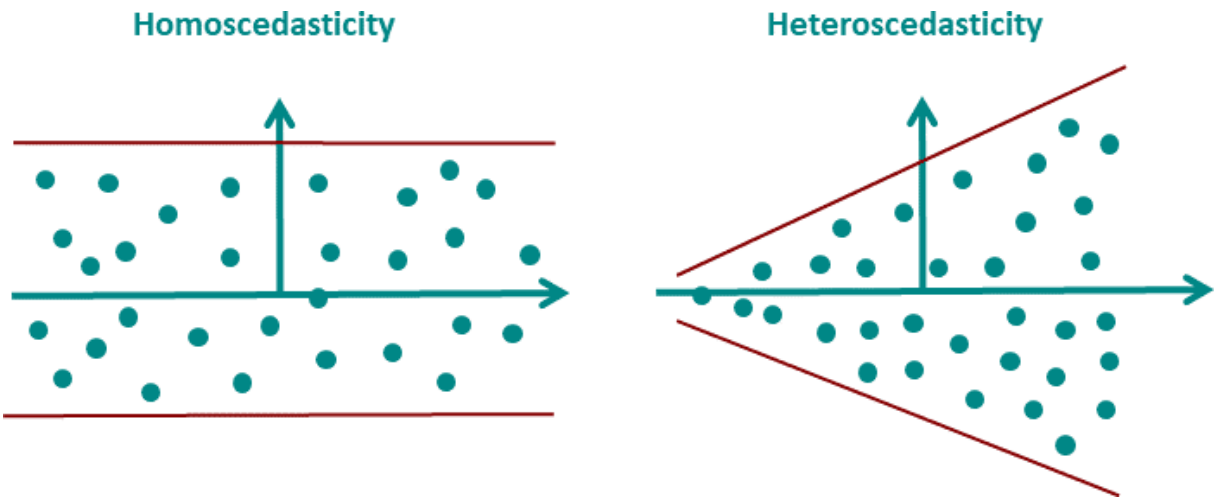
In the upper left graph, there is a linear relationship between the dependent and the independent variable, hence the regression line can be meaningfully put in. In the right graph you can see that there is a clearly non-linear relationship between the dependent and the independent variable. Therefore, it is not possible to put the regression line through the points in a meaningful way. For that reason, the coefficients cannot be meaningfully

interpreted by the regression model and there could be errors in the prediction that are greater than thought.

Therefore it is important to check beforehand whether a linear relationship between the dependent variable and each of the independent variables exists. This is usually checked graphically.

15.2.2.8 Homoscedasticity

Since in practice the regression model never exactly predicts the dependent variable, there is always an error. This very error must have a constant variance over the predicted range.



To test homoscedasticity, i.e. the constant variance of the residuals, the dependent variable is plotted on the x-axis and the error on the y-axis. Now the error should scatter evenly over the entire range. If this is the case, homoscedasticity is present. If this is not the case, heteroskedasticity is present. In the case of heteroscedasticity, the error has different variances, depending on the value range of the dependent variable.

15.2.2.9 Normal distribution of the error

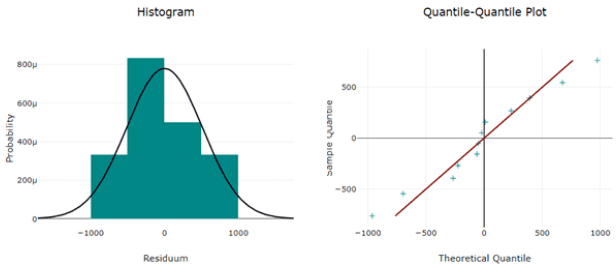
The next requirement of linear regression is that the error epsilon must be normally distributed. There are two ways to find it out: One is the analytical way and the other is the graphical way. In the analytical way, you can use either the Kolmogorov-Smirnov test or the Shapiro-Wilk test. If the p-value is greater than 0.05, there is no deviation of the data from the normal distribution and one can assume that the data are normally distributed.

Analytical

Copy

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistics	df	p-value	Statistics	df	p-value
Residuum	0,16	12	0,873	0,973	12	0,936

Graphically



However, these analytical tests are used less and less because they tend to attest normal distribution for small samples and become significant very quickly for large samples, thus rejecting the null hypothesis that the data are normally distributed. Therefore, the graphical variant is increasingly used.

In the graphical variant, either the histogram is looked at or, even better, the so-called QQ-plot or Quantile-Quantile-plot. The more the data lie on the line, the better the normal distribution.

15.2.2.10 Multicollinearity

Multicollinearity means that two or more independent variables are strongly correlated with one another. The problem with multicollinearity is that the effects of each independent variable cannot be clearly separated from one another.

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$\hat{x}_1 = b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$\hat{x}_2 = b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a$$

$$\hat{x}_k = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + a$$

Toleranz

$$T = 1 - R^2$$

Coefficient of determination

Warning:

$$T < 0.1$$

VIF (Variance Inflation Factor)

$$VIF = \frac{1}{1 - R^2}$$

Coefficient of determination

Warning:

$$VIF > 10$$

If, for example, there is a high correlation between x_1 and x_2, then it is difficult to determine b_1 and b_2. If both are e.g. completely equal, the regression model does not know how large b_1 and b_2 should be, becoming unstable.

This is of course not tragic if the regression model is only used for a prediction; in the case of a prediction, one is only interested in the prediction, but not in how great the influence of the respective variables is. However, if the regression model is used to measure the influence of the independent variables on the dependent variable, and if multicollinearity exists, the coefficients cannot be interpreted meaningfully.

15.2.2.11 Significance test and Regression

The regression analysis is often carried out in order to make statements about the population based on a sample. Therefore, the regression coefficients are calculated using the data from the sample. To rule out the possibility that the regression coefficients are not just random and have completely different values in another sample, the results are statistically tested with significance test. This test takes place at two levels.

- Significance test for the whole regression model
- Significance test for the regression coefficients

It should be noted, however, that the assumptions in the previous section must be met.

Significance test for the regression model

Here it is checked whether the coefficient of determination R2 in the population differs from zero. The null hypothesis is therefore that the coefficient of determination R2 in the population is zero. To confirm or reject the null hypothesis, the following F-test is calculated

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k}$$

The calculated F-value must now be compared with the critical F-value. If the calculated F-value is greater than the critical F-value, the null hypothesis is rejected and the R2 deviates from zero in the population. The critical F-value can be read from the F-distribution table. The denominator degrees of freedom are k and the numerator degrees of freedom are n-k-1.

Significance test for the regression coefficients

The next step is to check which variables have a significant contribution to the prediction of the dependent variable. This is done by checking whether the slopes (regression coefficients) also differ from zero in the population. The following test statistics are calculated in order to analyze it

Significance test for the regression coefficients

The next step is to check which variables have a significant contribution to the prediction of the dependent variable. This is done by checking whether the slopes (regression coefficients) also differ from zero in the population. The following test statistics are calculated in order to analyze it

$$t = \frac{b_j}{s_{b_j}}$$

where bj is the jth regression coefficient and sb_j is the standard error of bj. This test statistic is t-distributed with the degrees of freedom n-k-1. The critical t-value can be read from the t-distribution table.

15.2.2.12 Example linear regression

As an example of linear regression, a model is set up to predict the body weight of a person. The dependent variable is therefore the body weight, and the body height, age and gender are selected as independent variables. The following fictitious example data set is available:

Weight	**Height**	**Age**	**Gender**
79	1.80	35	Male
69	1.68	39	Male
73	1.82	25	Male
95	1.70	60	Male
82	1.87	27	Male
55	1.55	18	Female
69	1.50	89	Female
71	1.78	42	Female
64	1.67	16	Female
69	1.64	52	Female

This is how it works with DATAtab:

After you have copied the data into the statistics calculator, you must select the relevant variables. After that you will get the results in table form. The tables are shown below.

Table 3: Results of the linear regression

Model Summary

Copy Settings

R	R^2	Adjusted R^2	Standard error of the estimate
0.87	0.75	0.63	6.59

ANOVA

Copy Settings

Model	df	F	p
Regression	3	6.12	.012

Coefficients

Copy Settings

	Unstandardized Coefficients	Standardized Coefficients				95% confidence interval for B	
Model	B	Beta	Standard error	t	p	lower bound	upper bound
(Constant)	-24.41		47.64	-0.51	.627	-140.99	92.17
Height	47.38	0.52	27.63	1.71	.137	-20.23	114.99
Age	0.3	0.61	0.11	2.6	.041	0.02	0.58
Male	8.92	0.43	5.6	1.59	.162	-4.79	22.63

15.2.2.13 Interpretation of the results

It can be seen from the table that 75.4% of the **variation in** weight can be determined by height, age and gender. On average, the model overestimates by 6.587 in predicting the weight of a person. The **regression equation** is as follows:

Weight = 47.379 x **Height** + 0.297 x **Age** + 8.922 x **is_male** -24.41.

If, for example, the age increases by one year, the weight increases by 0.297 kg according to the model. In the case of the dichotomous variable gender, the slope is to be interpreted as the difference. According to the model, a man weighs 8.922 kg more than a woman. If all independent variables are zero, the result is a weight of -24.41, thus the constant (intercept).

The **standardized coefficients beta** are independent of the measured variable and always lie between -1 and 1. The larger the amount of beta, the greater the contribution of the respective independent variable to the elucidation of the variance of the dependent variable. In this regression analysis, the variable age has the greatest influence on the variable weight.

The calculated coefficients refer to the **sample** used for the calculation of the **regression analysis**. Therefore, it is of interest whether the **B-values** only deviate from zero by chance or whether this is also the case in the **population**. For this purpose, the null hypothesis is made that the respective calculated B-value is equal to zero in the population. If this is the case, it means that the respective independent variable has no significant influence on the dependent variable.

The sig. value indicates whether a variable has a **significant** influence. Sig. values smaller than 0.05 are considered significant. In this example, only age can be regarded as significant.

15.2.2.14 Presenting the results of the regression

When presenting your results, you should include the estimated effect, that is, the regression coefficient, the standard error of the estimate, and the p-value. Of course, it is also useful to interpret the regression results so that everyone knows what the regression coefficients mean.

For example: a significant relationship ($p < .041$) was found between a person's weight and a person's age.

If a simple linear regression was calculated, the result can also be displayed using a scatter plot.

15.3 Logistic Regression

Logistic regression is a **special case of regression analysis** and is calculated when the **dependent variable** is **nominally scaled or ordinally scaled.** This is the case, for example, with the variable "purchase decision" with two characteristics, "buys a product" and "does not buy a product".

Logistic regression analysis thus represents the counterpart to linear regression, in which the dependent variable of the regression model must at least be interval scaled.

With logistic regression it is possible to explain the dependent variable or to estimate the probability of occurrence of the manifestations of the variable.

Logistic regression is used in various fields and some different contextual examples follows below.

- First, an example from **business administration**: For an online retailer, you have to predict which product a certain customer would be most likely to buy. For this purpose, you receive a data set with the visitors of the website and the sales of the online retailer.

- Second, an example from **medicine**: You want to investigate whether a person is "susceptible" or "not susceptible" to a certain disease. For this purpose, you receive a data set with diseased and non-diseased persons as well as other medical parameters.

- Finally, an example from **political science**: In a population survey, you want to investigate whether a person would "vote" or "not vote" for party A if there was an election?

15.3.1 What is logistic regression?

In the basic form of logistic regression, **dichotomous variables (0 or 1)** can be predicted. For this purpose, the probability for the occurrence of **characteristic 1 (=characteristic present) is** estimated.

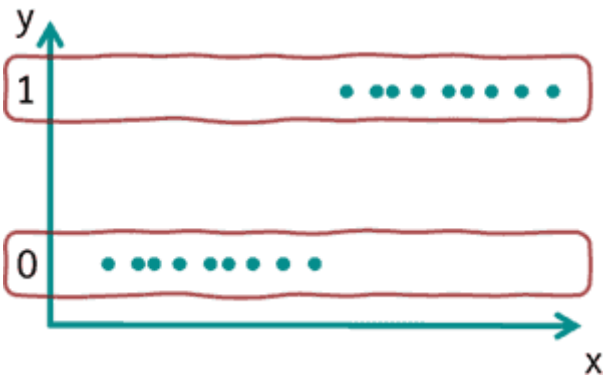


Figure 106: Dichotomous variables in logistic regression

In medicine, for example, it is often necessary to find out which variables have an influence on the occurrence of a disease. In this case, 0 could stand for "not diseased" and 1 for "diseased". Subsequently, the influence of age, gender, and smoking status (smoker or not) on this disease could be examined.

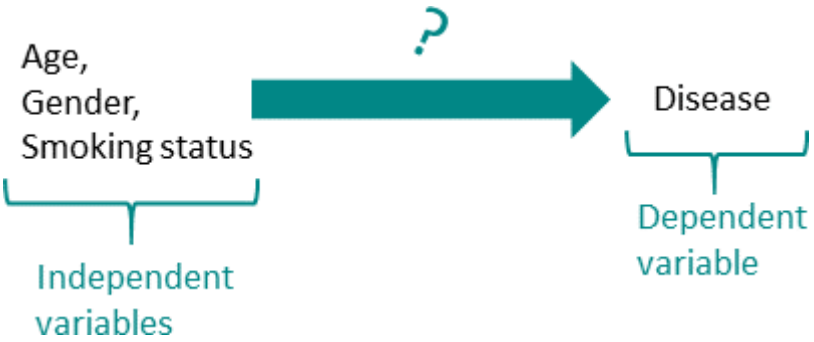


Figure 107: Factors influencing a disease in the regression model

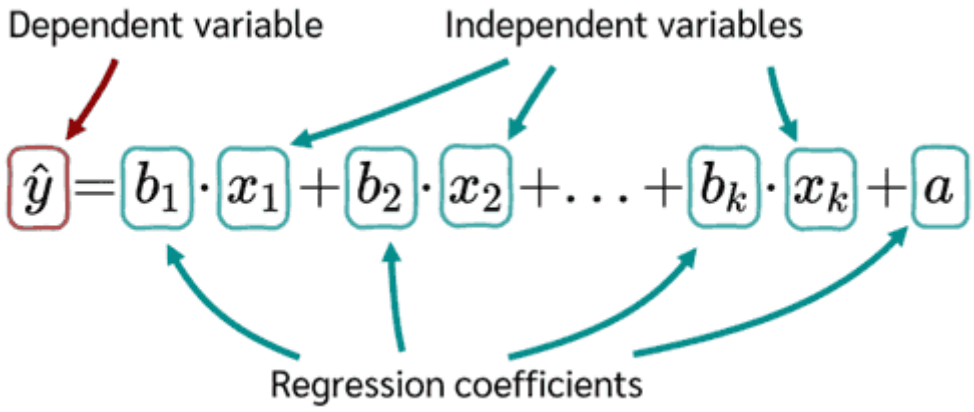
15.3.2 Logistic regression and probabilities

In linear regression, the independent variables (e.g., age and gender) are used to estimate the specific value of the dependent variable (e.g., body weight).

In logistic regression, on the other hand, the dependent variable is dichotomous (0 or 1) and the probability that expression 1 occurs is estimated. Returning to the example above, this means: How likely is it that the disease is present if the person under consideration has a certain age, sex and smoking status.

15.3.3 Calculate logistic regression

To build a logistic regression model, the linear regression equation is used as the starting point.



However, if a linear regression were simply calculated for solving a logistic regression, the following result would appear graphically:

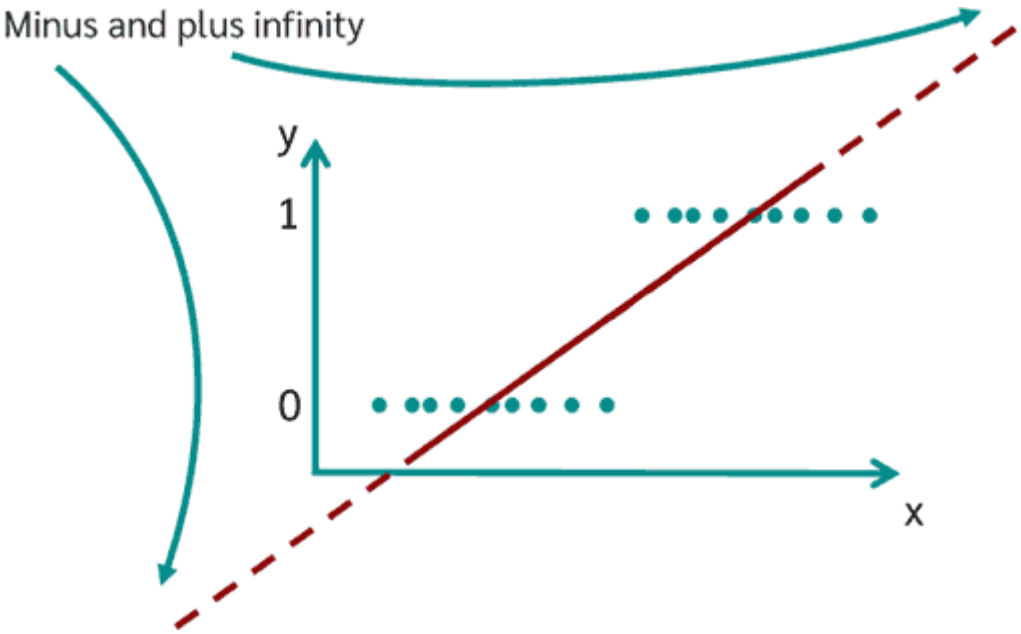


Figure 108: Limits of linear regression

As can be seen in the graph, however, **values between plus and minus infinity** can now occur. The goal of logistic regression, however, is to estimate the probability of occurrence and not the value of the variable itself. Therefore, the equation must still be transformed.

To do this, it is necessary to restrict the value range for the prediction to the range between 0 and 1. To ensure that only values between 0 and 1 are possible, the **logistic function f** is used.

15.3.4 Logistic function

The logistic model is based on the logical function. The special thing about the logistic function is that for values between minus and plus infinity, it always assumes only values between 0 and 1.

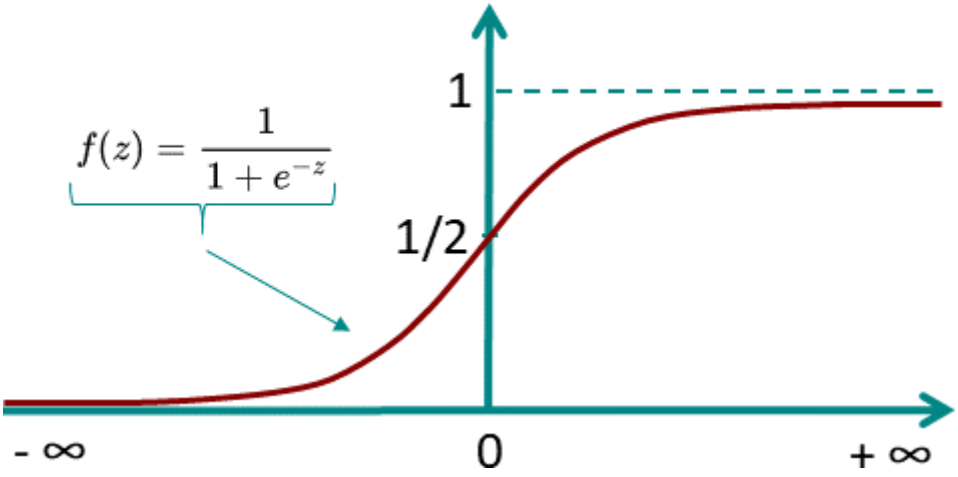


Figure 109: The logistic function

So, the logistic function is perfect to describe the **probability P(y=1)**. If the logistic function is now applied to the upper regression equation the result is:

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

This now ensures that, no matter in which range the *x-values are*, only numbers between 0 and 1 come out. The new graph now looks like this:

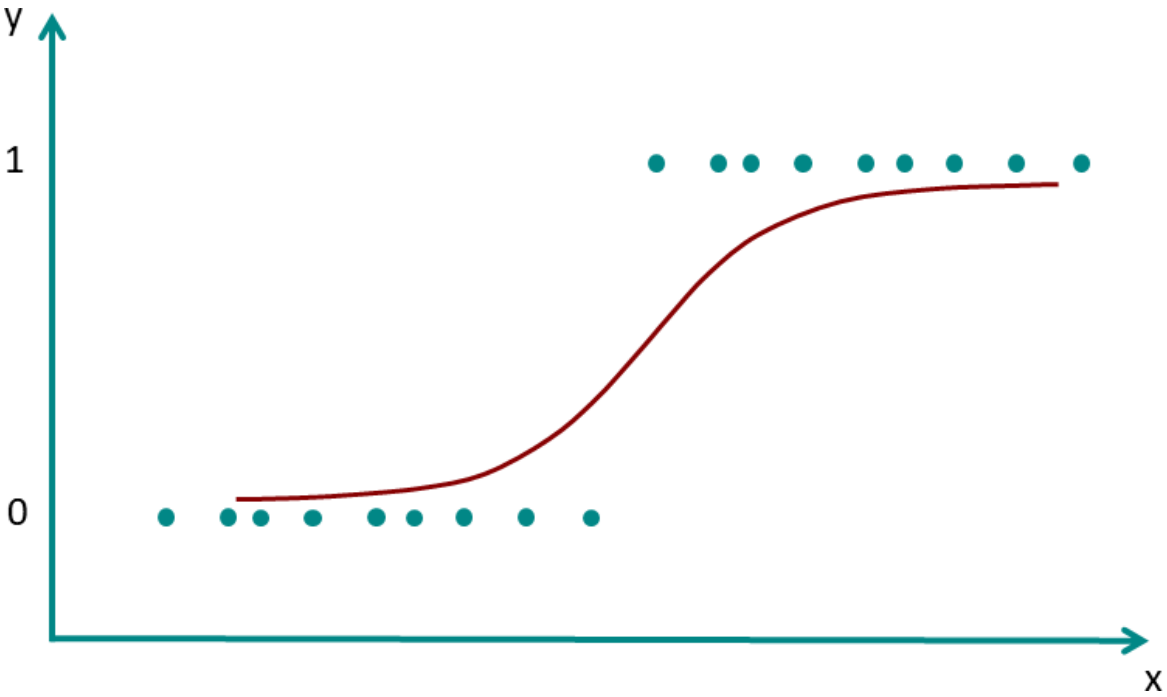


Figure 110: Approximation of the logistic function

The probability that for given values of the independent variable the dichotomous dependent variable y is 0 or 1 is given by:

$$P(y = 1|x_1, \ldots, x_n) = \frac{1}{1 + e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

$$P(y = 0|x_1, \ldots, x_n) = 1 - \frac{1}{1 + e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

.

To calculate the probability of a person being sick or not using the logistic regression for the example above, the model parameters b1, b2, b3 and a must first be determined. Once these have been determined, the equation for the example above is:

$$P(\text{Diseased}) = \frac{1}{1 + e^{-(b_1 \text{Age} + b_2 \text{Gender} + b_3 \text{Smoking status} + a)}}$$

15.3.5 Maximum Likelihood Method

To determine the model parameters for the **logistic regression equation**, the **Maximum Likelihood Method** is applied.

The maximum likelihood method is one of several methods used in statistics to estimate the parameters of a mathematical model.

Another well-known estimator is the least squares method, which is used in linear regression.

15.3.5.1 The Likelihood Function

To understand the **maximum likelihood method**, we introduce the **likelihood function** *L*. *L* is a function of the unknown parameters in the model. In the case of logistic regression, these are b_1 ,... b_n , a. Therefore, we can also write $L(b_1$,... b_n , $a)$ or $L(\vartheta)$ if the parameters are summed in θ.

L(θ) now indicates how probable it is that the observed data occur. With the change of θ, the probability that the data will occur as observed changes.

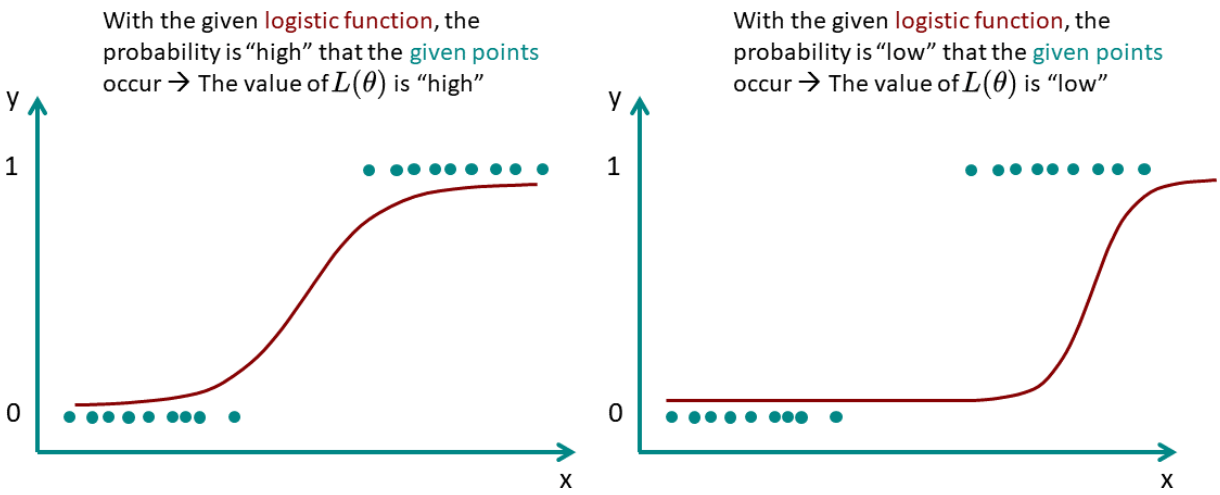

With the given logistic function, the probability is "high" that the given points occur → The value of $L(\theta)$ is "high"
y
1
0
x
With the given logistic function, the probability is "low" that the given points occur → The value of $L(\theta)$ is "low"
y
1
0
x

Figure 111: Likelihood function

15.3.5.2 Maximum Likelihood Estimator

The **maximum likelihood estimator** can be applied in the estimation of complex nonlinear as well as linear models. In the case of logistic regression, the goal is to find the parameters b_1 ,... b_n , *a that* maximize the so-called **log-likelihood function** *LL(ϑ).*

The log-likelihood function is simply the logarithm of *L(ϑ)*.

Various algorithms have been established over the years for this nonlinear optimization, such as the **stochastic gradient descent method**.

15.3.6 Multinomial logistic regression

As long as the dependent variable has two characteristics (e.g. *male*, *female*), i.e. is dichotomous, **binary logistic regression** is used. However, if the dependent variable has more than two instances, e.g. which mobility concept describes a person's journey to work (*car*, *public transport*, *bicycle*), **multinomial logistic regression** must be used.

Each expression of the mobility variable (*car*, *public transport*, *bicycle*) is transformed into a new variable. The one variable mobility concept becomes the three new variables:

- *car is used*
- *public transport is used*
- *bicycle is used*

Each of these new variables then only has the two values *yes* or *no*, e.g. the variable *car is used* only has the two answer options *yes* or *no* (either it is used or not).

Thus, for the one variable "mobility concept" with three values, there are three new variables with two values each: *yes* and *no* (*0* and *1*). Three logistic regression models are now created for these three variables.

15.3.7 Interpretation of the results

The relationship between the dependent and independent variable in logistic regression is not linear, hence the regression coefficients cannot be interpreted in the same way. For this reason, **odds ratios** are interpreted in **logistic regression**.

- **Linear Regression:**

  An independent variable is said to be "good" or "fit" if it is strongly correlated with the dependent variable.

- **Logistic Regression:**

An independent variable is called "good" or "suitable" if it makes it possible to distinguish significantly between the groups of the dependent variable.

15.3.8 Odds Ratios

An **odds ratio (OR)** is a statistical measure used to determine the strength of association or effect size between two events or groups, often in case-control studies. It compares the odds of an event occurring in one group to the odds of it occurring in another group.

Odds represent the ratio of the probability of an event happening to it not happening. Odds ratio is the ratio of these odds between two groups.

A detailed explanation of Odds Ratios in Logistic Regression can be found in the section “Odds Ratios in Logistic Regression” at the end of this chapter.

15.3.9 Pseudo-R squared

In a linear regression, the coefficient of determination R2 indicates the proportion of the explained variance. In logistic regression, the dependent variable is scaled nominally or ordinally and it is not possible to calculate a variance, so the coefficient of determination cannot be calculated in logical regression.

However, in order to make a statement about the quality of the logistic regression model, so-called pseudo coefficients of determination have been established, also called pseudo-R squared. Pseudo coefficients of determination are constructed in such a way that they lie between 0 and 1 just like the original coefficient of **determination**. The best known coefficients of determination are the **Cox and Snell R-square** and the **Nagelkerke R-square**.

15.3.10 Null Model

For the calculation of the Cox and Snell R-square and the Nagelkerke R-square, the likelihood from the so-called null model L0 and the likelihood L1 from the calculated model (full model) is needed. The null model is a model in which no independent variables are included, L1 is the likelihood of the model with the dependent variables.

15.3.11 Cox and Snell R-square

In the Cox and Snell R-square, the ratio of the likelihood function of the null model L0 and L1 is compared. The better the model being fitted (full model) is compared to the null model, the lower the ratio between L0 and L1. The Cox and Snell R-square is obtained with:

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_1}\right)^{2/n}$$

15.3.12 Nagelkerkes R-square

The Cox and Snell pseudo-determination measure cannot become 1 even with a model with a perfect prediction, this is corrected with the R-square of Nagelkerkes. The Nagelkerkes pseudo coefficient of determination becomes 1 if the model being fitted gives a perfect prediction with a probability of 1.

$$R_N^2 = \frac{1 - \left(\frac{L_0}{L_1}\right)^{2/n}}{1 - L_0^{2/n}}$$

15.3.13 McFadden's R-square

The McFadden's R-square also uses the null model and the model being fitted to calculate the R2.

$$R_{McF}^2 = 1 - \frac{ln(L_1)}{ln(L_0)}$$

15.3.14 Chi2 Test and Logistic Regression

In the case of logistic regression, the Chi-square test tells whether the model is overall significant or not.

Chi-Squared Test

Copy Word Copy Excel

Chi2	df	p
8.79	3	.032

Here two models are compared. In one model all independent variables are used and in the other model the independent variables are not used.

Model 1:
With independent variables

$$\frac{1}{1+e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

Model 2:
Without independent variables

$$\frac{1}{1+e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

Now the Chi-square test compares how good the prediction is when the dependent variables are used and how good it is when the dependent variables are not used.

The Chi-square test now tells us if there is a significant difference between these two results. The null hypothesis is that both models are the same. If the p-value is less than 0.05, this null hypothesis is rejected.

15.3.15 Example logistic regression

As an example for the logistic regression, the purchasing behaviour in an online shop is examined. The aim is to determine the influencing factors that lead a person to buy immediately, at a later time or not at all from the online shop after visiting the website. The online shop provides the data collected for this purpose. The dependent variable therefore has the following three characteristics:

- Buy now
- Buy later
- Don't buy

Gender, age and time spent in the online shop are available as independent variables.

Purchasing behaviour	Gender	Age	Time spent in online shop
Buy now	female	22	40
Buy now	female	25	78
Buy now	male	18	65
...	...	...	...
Buy later	female	27	28
Buy later	female	27	15
Buy later	male	48	110
...	...	...	...
Don't buy	female	33	65
Don't buy	female	43	34

15.3.16 Calculating logistic regression with DATAtab

Logistic regressions, similar to linear regression models, can be easily and quickly calculated with DATAtab.

If you want to recalculate the example above, simply copy and paste simply copy the table on purchasing behavior in the online store into DATAtab's statistics calculator.

Then select the *Regression* tab and click on the desired variables. You directly get the results below in table form.

Result

Copy Word Copy Excel

Total number of cases	Correct assignments	In percent
24	17	70.83 %

Classification table

Copy Word Copy Excel

		Predicted		
		not Buy now	Buy now	Correct
Observed	not Buy now	13	3	81.25 %
	Buy now	4	4	50 %
	Total			70.83 %

Chi-Squared Test

Copy Word Copy Excel

Chi2	df	p
8.21	3	.042

Model Summary

Copy Word Copy Excel

-2 Log-Likelihood	Cox & Snell R^2	Nagelkerke R^2	McFadden's R^2
22.34	0.29	0.4	0.27

Model

Copy Word Copy Excel

	Coefficient B	Standard error	z	p	Odds Ratio	95% conf. interval
female	-1.53	1.2	1.28	.201	0.22	0.02 - 2.27
Age	-0.11	0.06	1.81	.07	0.9	0.8 - 1.01
Time spent in online shop	-0.02	0.03	0.53	.596	0.98	0.93 - 1.04
Constant	4.28	2.23	1.92	.055		

15.3.17 Odds Ratios in Logistic Regression

In this section we start with a quick overview of logistic regression, then dive into what odds are and how they work. From there, we'll break down the odds ratio, and finally, we'll bring it all together to see what the odds ratio means within the context of logistic regression.

15.3.18 Basic concept of logistic regression

As already discussed, in a regression analysis you want to infer or predict an outcome variable based on one or more other variables. The outcome variable is also called the dependent variable, and the other variables are independent variables.

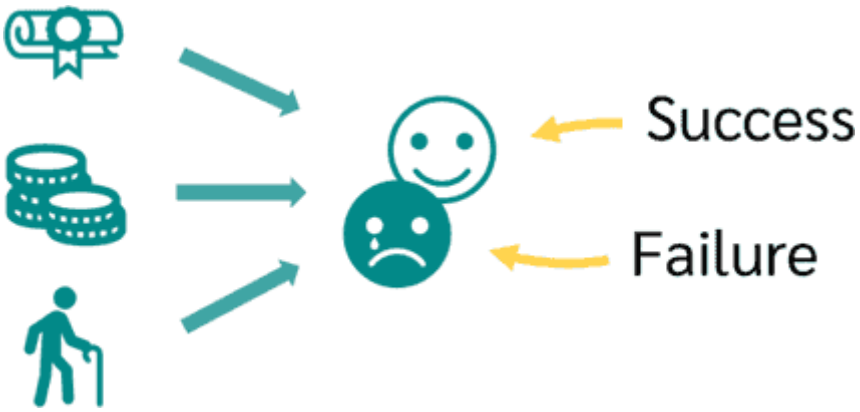
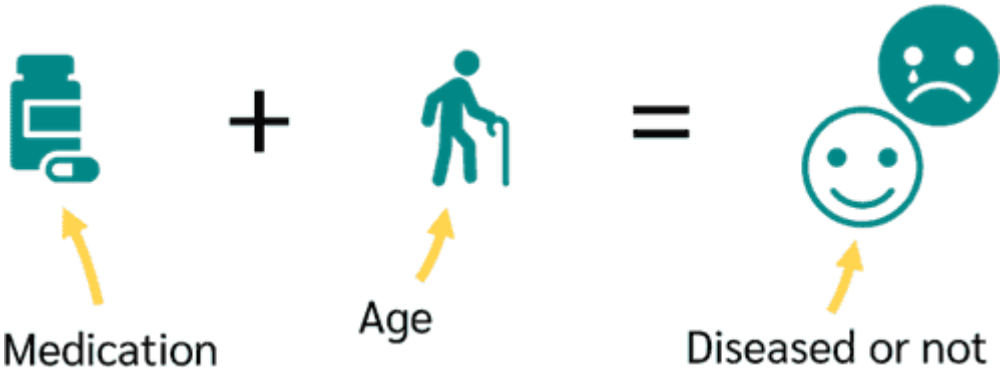


Figure 112: Example binary logistic regression

A binary Logistic Regression is a type of regression analysis used when the outcome variable is binary, meaning it has two possible values, like 'Yes' or 'No,' 'Success' or 'Failure.'

15.3.19 Example binary logistic regression

Let's say we are researchers, and we want to know whether a particular medication and a person's age have an influence on whether a person gets a certain disease or not.



So, the outcome we're interested in is whether the patients developed the disease or did not develop it. And our independent variables are medication and age of a person. Now with the help of a logistic Regression, we want to infer or predict the outcome variable based on the independent variables. Now let’s take a look at what odds mean.

15.3.20 Odds in logistic regression

Let’s say we have two possible outcomes of something: success and failure. For example, if a therapy is successful or not. The probability that the therapy is successful is 0.7 (or 70%) and thus the probability of failure is 1 - 0.7 = 0.3.

15.3.21 What are the odds?

Odds are defined as the ratio of the probability of success and the probability of failure. In other words, odds represent the ratio of the probability of an event happening to the probability of it not happening.

$$\frac{\text{Success}}{\text{Failure}}$$

If we look at our example, the odds are 0.7 divided by 0.3, which equals 2.33. This means the event “success” is 2.33 times more likely to happen than not.

$$\text{Odds} = \frac{\checkmark\ \text{Success}}{\times\ \text{Failure}} = \frac{0.7}{0.3}$$

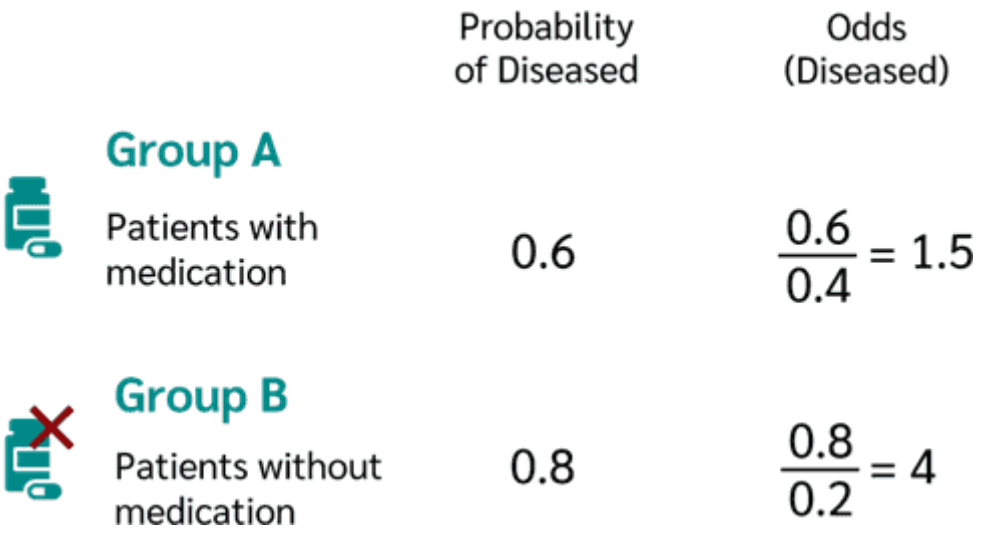
So odds give us a measure of the likelihood of an event happening versus not happening.

$$\text{Odds} = \frac{\text{Probability of event } \mathbf{happening}}{\text{Probability of it } \mathbf{not\ happening}}$$

15.3.22 What are Odds Ratios?

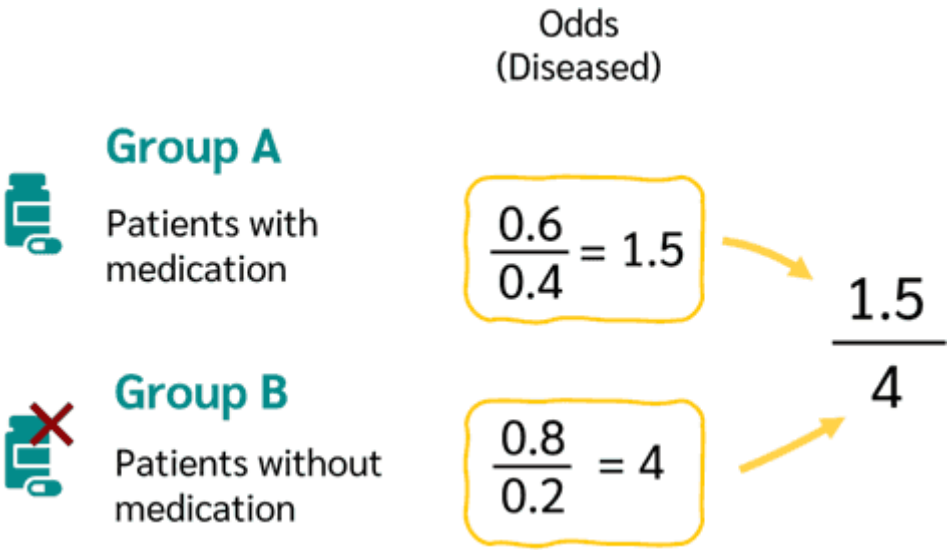
Let's say we have a Group A (Patients with medication) and a Group B (Patients without medication). In Group A, we calculated a probability of 60% (or 0.6) of getting diseased. So the odds of getting diseased is 0.6 divided by 0.4, which is 1.5.

In Group B, where the patients didn’t get the medication, the probability of getting diseased is 80% (or 0.8). So the odds in Group B of getting diseased are 0.8 divided by 0.2, which is 4.

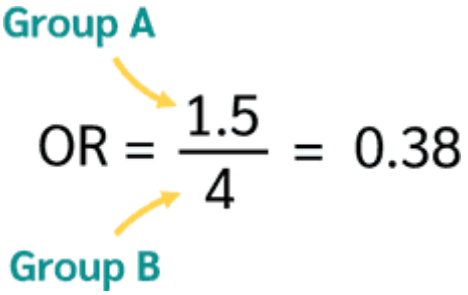


With the odds ratio, we can now compare the two groups. To do this, we can compare the odds of getting the disease in Group A relative to the odds of getting the disease in Group B.

The odds ratio is simply calculated by dividing the odds in Group A by the odds in Group B. This results in an odds ratio of 0.38.



The odds ratio of 0.38 means that the odds of being diseased in Group A are 0.38 times the odds of being diseased in Group B.



Of course we can also switch the order, then the odds ratio would be the odds in Group B divided by the odds in Group A. In this case, the odds ratio of approximately 2.67 means that the odds of being diseased in Group B are 2.67 times higher than the odds of being diseased in Group A.

If the odds ratio is greater than 1, the event is more likely to occur in the first group. If it's less than 1, the event is less likely in the first group.

15.3.23 Odds Ratio in Logistic Regression

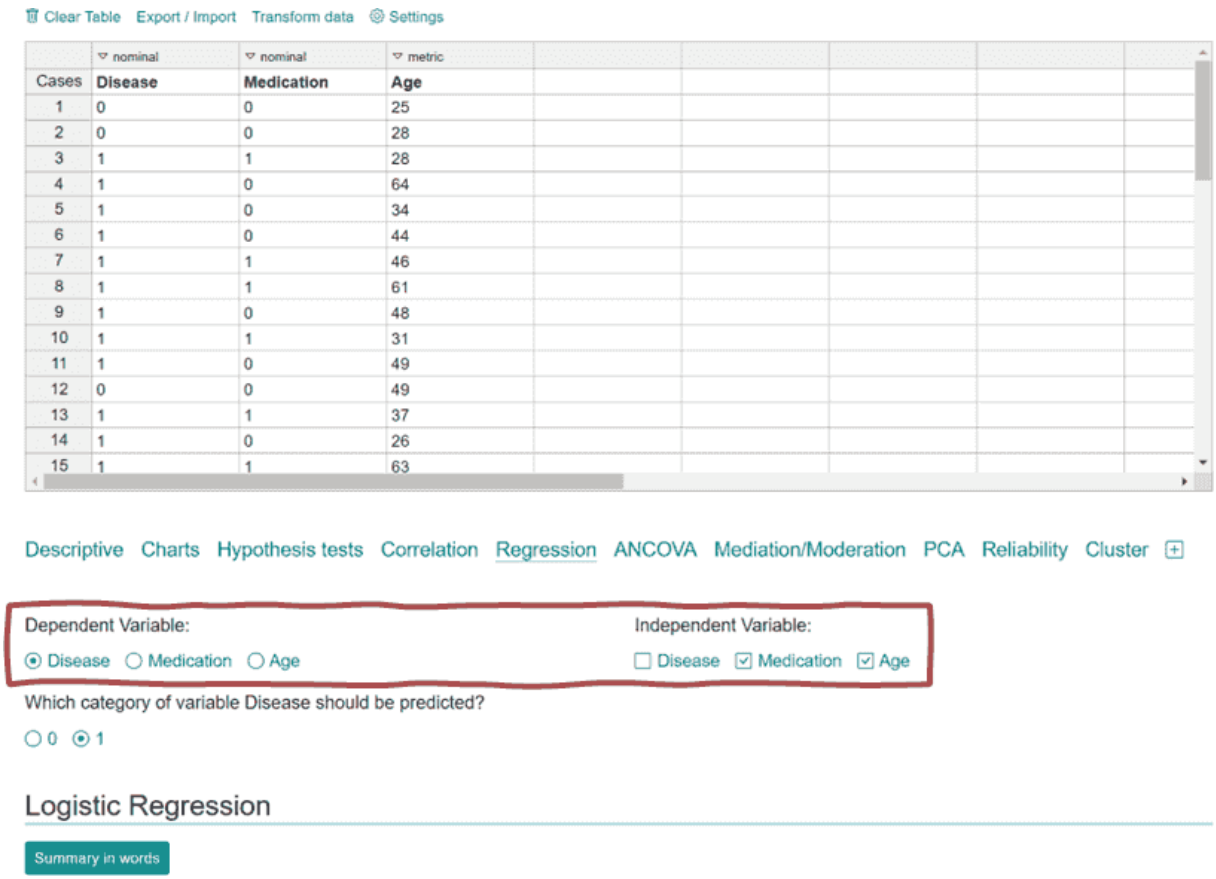
First of all, to calculate a logistic regression we need data. Let's say we have data from 50 patients. Our outcome variable is Disease, which is coded as 0 for 'not diseased' and 1 for 'diseased.' And we have two independent variables: Medication and Age. For the Medication variable, 0 indicates 'no medication,' and 1 indicates 'medication taken.'

Medic.	Age	Disease
0	25	0
0	28	0
1	28	1
0	64	1
0	34	1
0	44	1
1	46	1
...	...	...
...	...	...
...	...	...
1	25	0

15.3.24 Calculate Odds Ratios with DATAtab

Now we can use this data to calculate a logistic regression. When you click on this link, the data is directly opened in DATAtab. We want to calculate a logistic regression, so we choose Regression. Here we can select the dependent and the independent variable. So, we select “disease” as our dependent variable and “medication” and “age” as our independent variables.

Now we get the results of the logistic regression.



Cases	nominal Disease	nominal Medication	metric Age
1	0	0	25
2	0	0	28
3	1	1	28
4	1	0	64
5	1	0	34
6	1	0	44
7	1	1	46
8	1	1	61
9	1	0	48
10	1	1	31
11	1	0	49
12	0	0	49
13	1	1	37
14	1	0	26
15	1	1	63

Now we get the results of the logistic regression. Here we see the table that we will now take a closer look at.

	Coefficient B	Standard error	z	p	Odds Ratio
Constant	-1.45	1.16	1.25	.211	0.23
Medication	-0.45	0.59	0.76	.447	0.64
Age	0.04	0.02	1.69	.09	1.04

In the first column, we can see the coefficients that define our model. These coefficients can be entered into the logistic regression formula.

	Coefficient B	Standard error	z	p	Odds Ratio
Constant	-1.45	1.16	1.25	.211	0.23
Medication	-0.45	0.59	0.76	.447	0.64
Age	0.04	0.02	1.69	.09	1.04

$$P = \frac{1}{1 + e^{-(-1.45 - 0.45 \cdot Medication + 0.04 \cdot Age)}}$$

Now, we just need to enter a value for Medication—such as 1, indicating the patient received medication—and a value for Age, for example, 50.

$$P = \frac{1}{1 + e^{-(-1.45 - 0.45 \cdot 1 + 0.04 \cdot 50)}} = 0.55$$

Medication Age

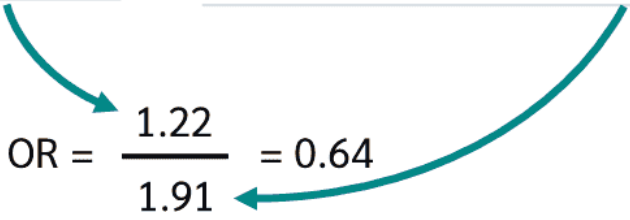
Then we can calculate the probability. In this case, the probability of being diseased is 0.55, or 55%. Okay, but we're not interested in the odds alone—we're interested in the odds ratio.

Again, the odds ratio is simply a comparison of the odds of an event occurring in two different groups.

Therefore, we just need to compare the odds of a person who took the medication with the odds of a person who did not take the medication. So to get the odds ratio, we just need to divide the odds of a person who took the medication by the odds of a person who did not take the medication. This results in an odds ratio of 0.64.

	Value
Medication	1
Age	50
Probability	0.55
Odds	0.55 / (1 - 0.55) = 1.22

	Value
Medication	0
Age	50
Probability	0.66
Odds	0.66 / (1 - 0.66) = 1.91



The odds ratio of 0.64 for Medication indicates that for individuals who took the medication, the odds of the outcome diseased are 0.64 times the odds for those who did not take the medication.

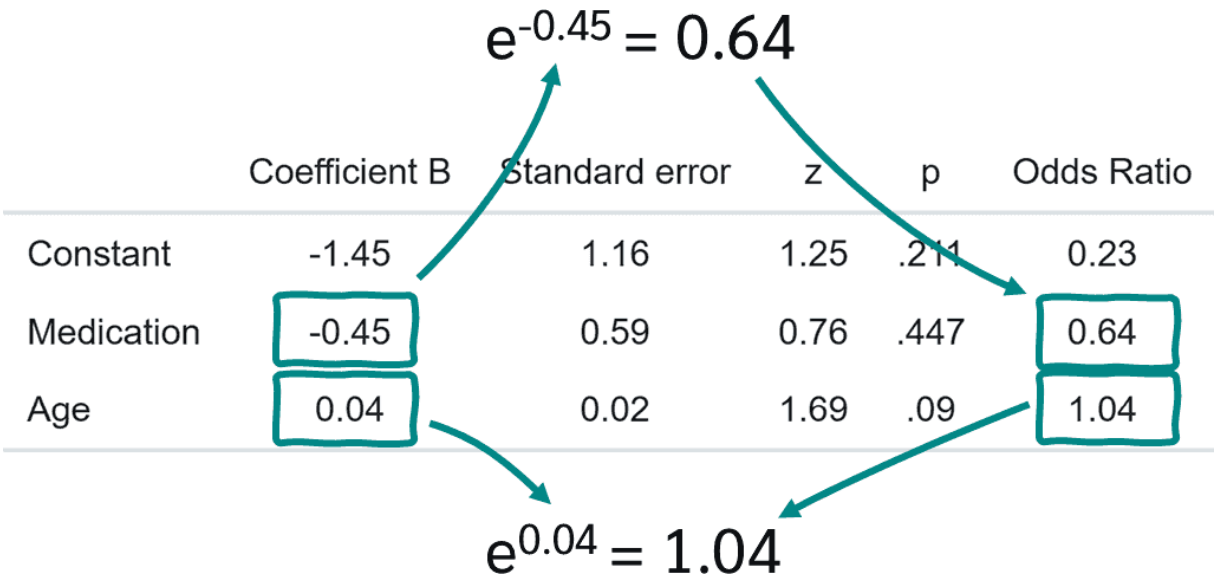
15.3.25 Odds ratios of continuous variables

With medication, we have two groups to compare. But what about a continuous variable like age? In this case, we simply look at what happens when we increase age by one unit. For example, we might compare the odds of the outcome for someone aged 50 versus someone aged 51. This allows us to calculate the odds ratio by comparing the two odds.

In this case, we get an odds ratio of 1.04. So for each one-year increase in age, the odds of the outcome 'diseased' increase by a factor of 1.04.

15.3.26 Odds ratio or exp(B)

There is one important thing: The odds ratio can actually be calculated simply by exponentiating each coefficient. So, exp(-0.45) is 0.64, which is the odds ratio of medication. And exp(0.04) is 1.04. which is the odds ratio for age.



16. Confidence Intervals

The confidence interval CI is the range in which a parameter (e.g. the mean value) lies with a certain probability.

16.1 Why do we need the confidence interval?

In statistics, parameters of the population are often estimated based on a sample, such as the mean or the variance. However, these are only estimates and the true value in the population will be somewhere around these estimates. It is very useful to define a range or interval where the true value is most likely to lie.

16.2 Interpretation of confidence interval

If several samples are taken from a population, it is very likely that each sample will have a different mean value. However, we want to know the mean of the population, not the mean of the sample. The confidence interval is the range in which the true mean of the population lies with a certain probability.

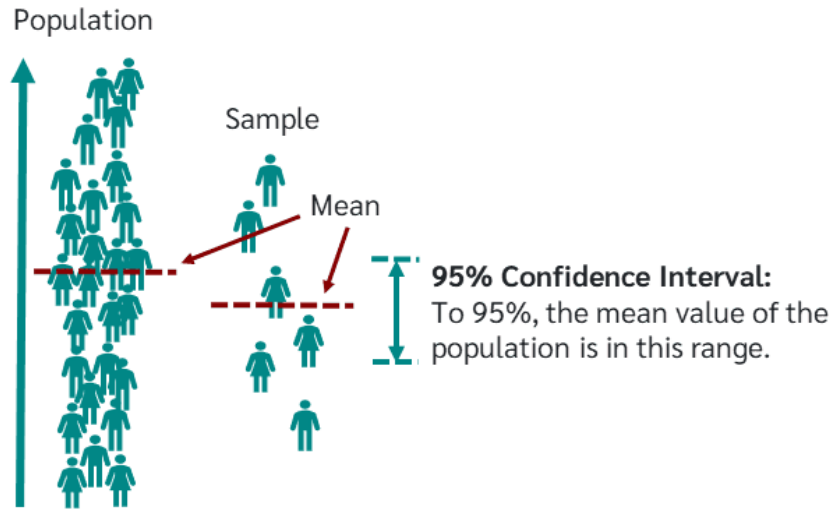


Figure 113: Confidence Interval

Caution! The above definition is widely used because it is easy to understand, but not all experts agree that it is correct. The following definition is definitely correct, but more complicated:

The 95% confidence interval (CI) is an interval calculated using sample data from an infinite series, 95% of which contain the population parameter. In the long run, 95% of such intervals contain the true mean.

So in fact there are two common ways to explain the confidence interval.

On the one hand, there's a simpler explanation of the confidence interval, but it's not correct when viewed from a frequentist statistics perspective. On the other hand, there's a slightly more complex explanation that is true.

16.2.1 Common Misinterpretation of the confidence interval

So, let's first address the simple but **incorrect** interpretation.

As already stated, this interpretation goes like this: "There is a 95% chance that the true parameter lies within a calculated confidence interval.“

What does this actually mean?

Imagine we have a population with a true mean value. This true mean value is the one we want to estimate. Although we don’t know this true mean, we can make an educated guess by taking a sample from the population.

From this sample, we calculate both the sample mean and the 95% confidence interval. The simplified interpretation is to say: The confidence interval provides a range within which the true mean lies with a certain probability. Or in case of the 95% confidence Interval, we would say: there's a 95% chance that the true value falls within this interval. However, this interpretation isn't accurate. But why?

In **frequentist statistics**, the true parameter is fixed and unchanging—it either falls within the calculated interval or it doesn't. The only thing that varies is the sample data we collect.

Since each new sample creates a new confidence interval, the "95%" refers to the long-run frequency: if you calculate many confidence intervals from repeated samples, 95% of them would contain the true parameter.

Once a specific interval is calculated, no probability can be assigned to the parameter being inside it.

So, for example in all these samples, the true value falls within the confidence interval, while in those two samples, it does not. In summary, you can't say, "There is a 95% chance that this interval contains the true parameter." because once the interval is calculated, it's either contains the parameter or it doesn't, and there's no probability left to assign in the frequentist sense.

Abbildung dazu geben!

16.2.2 Correct interpretation of the CI

Let's say we took a lot of random samples, and we calculated the mean value and the confidence interval of each sample.

The confidence interval can now be interpreted in the following way:

If we were to take an extremely large number of random samples and construct a confidence interval for each sample, 95% of those intervals would contain the true value, while 5% would not.

In other words, if we were to take 100 random samples, we would expect that, on average, 95 of the confidence intervals would contain the true value, while 5 would not.

Or the other way around: The confidence interval can be defined in terms of probability with respect to a single, theoretical sample that has yet to be realized.

Therefore, if you haven't drawn the sample yet, you can be 95% sure that the interval from the next sample you draw will contain the true value. But if you have taken the sample, the true value is either in the interval or not.

This means: Confidence Is About the Method, Not the Specific Interval. The "95% confidence" refers to the long-run reliability of the *method* you used to construct the interval. If you use this method repeatedly on different samples, you expect to capture the true parameter 95% of the time.

But once you've applied it and obtained a specific interval, you then cannot make a probability statement about whether this interval contains the fixed true parameter or not.

16.3 Calculate confidence interval

To calculate the confidence interval, the distribution function of the respective parameter (e.g. the mean value) in the population is required. Assuming that this distribution is normally distributed, the confidence interval for the mean is given by:

$$CI = \bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$$

Standard deviation

Mean value

Lower/Upper limit

z-value for the confidence level

Sample size

Where *x* is the sample mean, *n* is the sample size and *s* is the sample standard deviation. Plus and minus indicate the lower and upper limits of the confidence interval respectively.

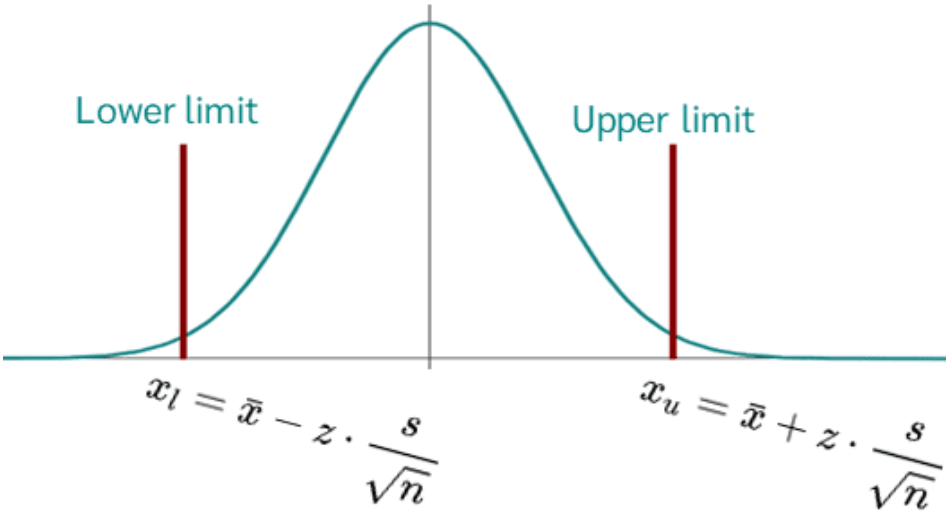


Figure 114: Calculating confidence intervals

If the sample is small, the t-distribution is used instead of the normal distribution. Then the *z* value is replaced by *t* and the formula is:

$$CI = \bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

16.4 Confidence interval 95%

To calculate the confidence interval, the probability that the population mean lies within the interval must be defined. The confidence level of 95% or 99% is very often used as probability. This probability is also called the **confidence coefficient**.

For the **95% confidence interval** and the **99% confidence interval**, the *z* values are as follows:

Confidence level	**95%**	**99%**
z-Value	1.96	2.58

If a 95% confidence interval is given, you can be 95% sure that the true value of the parameter lies within that interval.

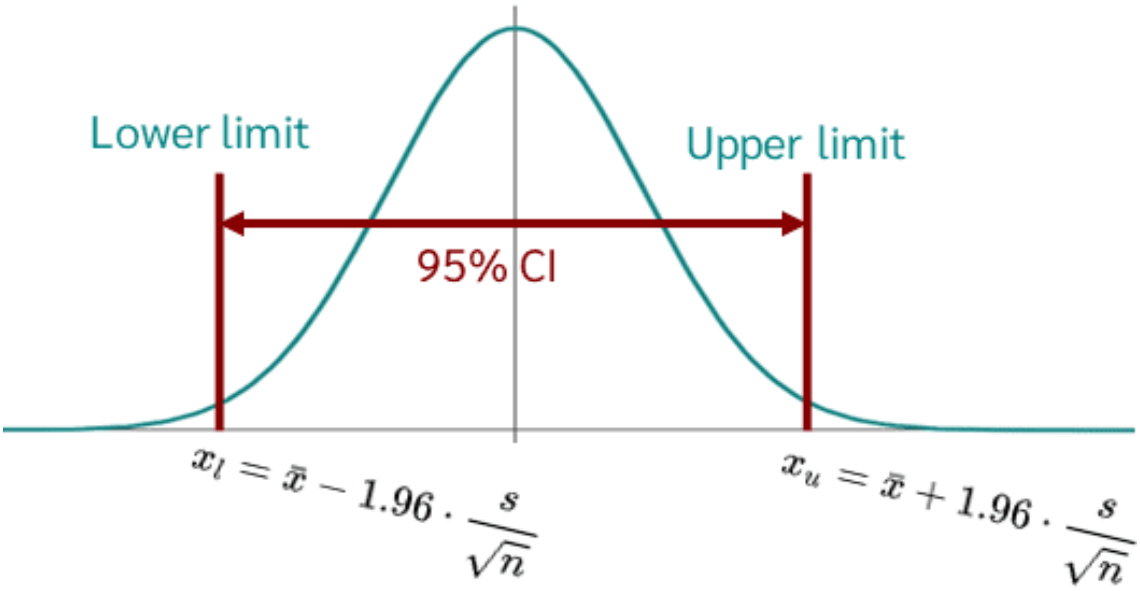


Figure 115: Calculating confidence intervals

16.5 Confidence interval for t-test

A t-test compares differences in means, e.g. you can use a t-test to test whether there is a difference in salary between men and women.

You actually want to make a statement whether there is a difference in salary in the population. Since you cannot survey the entire population, you use a sample. In this sample, there is a high probability of a difference in salary.

In order to be able to estimate approximately in which range the mean difference in the population lies, you calculate the confidence interval.

In the t-test calculator on DATAtab you can calculate the confidence interval of the mean difference.

17. Factor Analysis

Factor analysis is a method that aims to uncover structures in large variable sets. If you have a data set with many variables, it is possible that some of them are interrelated, i.e. correlate with each other. These correlations are the basis of factor analysis.

17.1 What is a factor?

In factor analysis, the factor can be seen as a hidden variable that influences several actually observed variables.

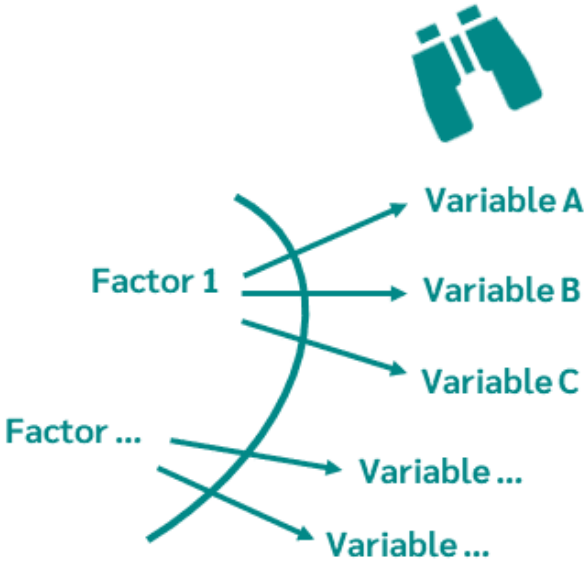


Figure 116: Factor Analysis

Or, in other words, several variables are observable phenomena of fewer underlying factors.

In factor analysis, therefore, the variables that are highly correlated with each other are combined. It is assumed that this correlation is due to a non-measurable variable, which is called a factor.

17.2 Example factor analysis

Factor analysis can be used to answer the following questions:

- What structure can be detected in the data?
- How can the data be reduced to some factors?

The following table contains examples of content that show where factor analysis is used in different fields of expertise.

Examples

	Question	Variable	Possible factors
Psychological	Can different personality traits be grouped into personality types?	Be sociable, be spontaneous, be curious, be nervous, be aggressive etc.	Neuroticism, Extraversion, Openness for new things, Conscientiousness ,Social compatibility
Business Administration	How can different cost types be summarized in cost characteristics?	Material costs, personnel costs, equipment costs, fixed costs etc.	Influenceability, urgency of coverage

17.3 Research questions factor analysis

A possible research question might be: Can different personality traits such as outgoing, curious, sociable, or helpful be grouped into personality types such as conscientious, extraverted, or agreeable?

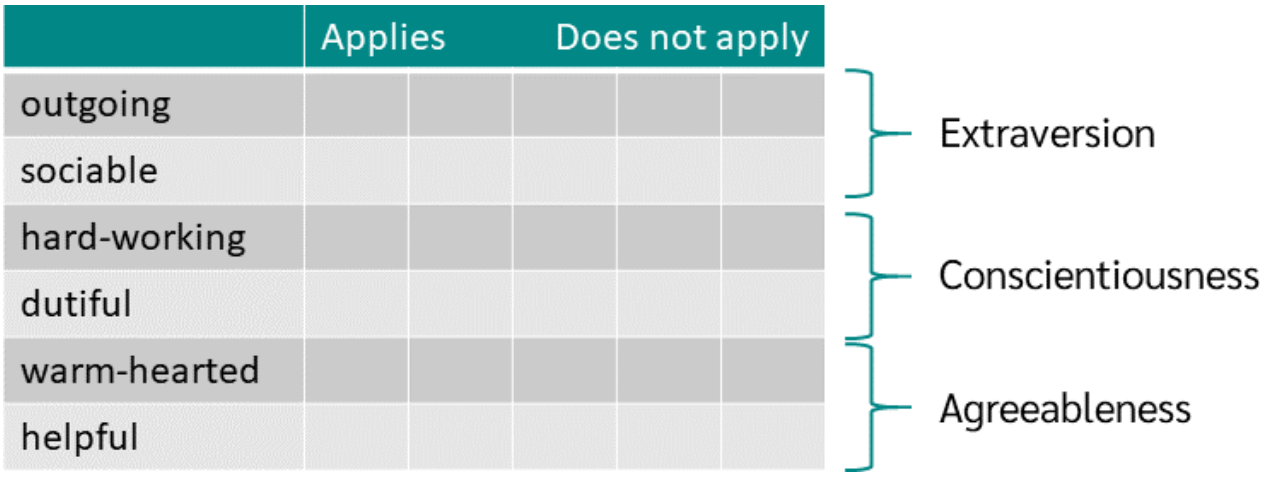


Figure 117: Basics of factor analysis

You want to find out whether some of the characteristics sociable, sociable, hard-working, conscientious, warm-hearted or helpful correlate with each other and can be described by an underlying factor. To find out, you created a small survey with DATAtab.

You have interviewed 20 people and have the results output to an Excel table. Here you can find the example data set for the Principal Component Analysis with which you can calculate the example directly online on DATAtab under Factor Analysis Calculator.

17.4 Factor load, eigenvalue, communalities

The important terms or characteristic values for a factor analysis are factor charge, eigenvalue and communalities. With their help, it is possible to see how strong the correlation between the individual variables and the factors is.

Factor load

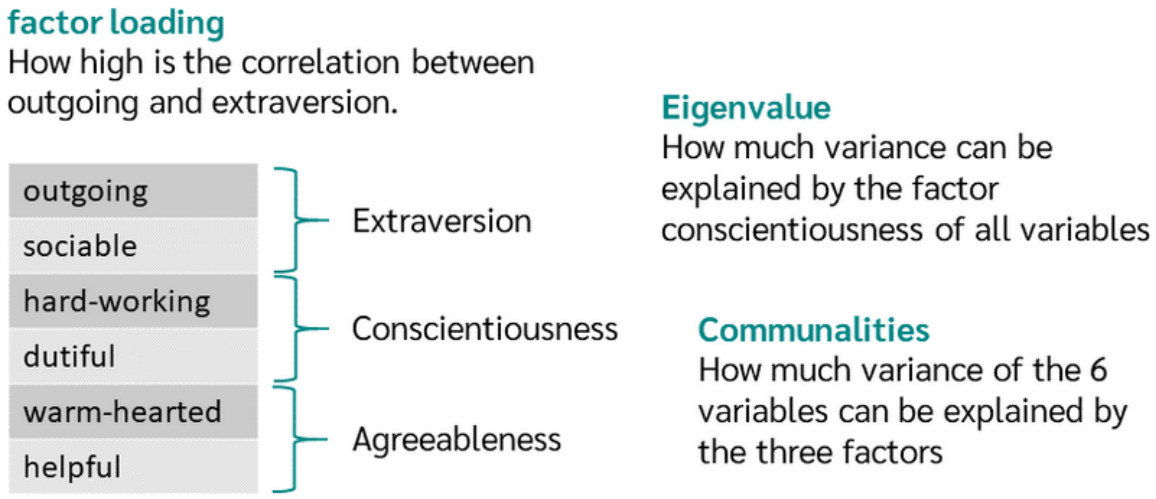
- Correlation between a variable and a factor
- Loading a variable to a factor

Eigenvalue

- The variance explained by a factor
- Sum of the squared factor charges

Communalities

- Variance of the variables, which is explained by all factors
- Sum of the squared factor charges of a variable



17.5 Correlation Matrix

The first step in factor analysis is to calculate the correlation matrix. Starting from the correlation matrix, the so-called eigenvalue problem is solved, which is used to calculate the factors.

Correlation matrix

Copy /

	outgoing	sociable	hard-working	dutiful	warm-hearted	helpful
outgoing	1	0.583	-0.132	-0.02	-0.038	-0.219
sociable	0.583	1	0.074	0.326	0.27	-0.257
hard-working	-0.132	0.074	1	0.343	0.002	0.03
dutiful	-0.02	0.326	0.343	1	0.001	-0.254
warm-hearted	-0.038	0.27	0.002	0.001	1	0.52
helpful	-0.219	-0.257	0.03	-0.254	0.52	1

17.6 Factor Analysis and dimensionality

It is important to note, however, that factor analysis does not give a "clear" answer as to how many factors must be used and how these factors can then be interpreted.

There are two common methods to determine the number of required factors: the eigenvalue criterion (Kaiser criterion) and the scree test.

17.6.1 Eigenvalue criterion (Kaiser criterion)

In order to determine the dimensions, i.e. the number of factors, with the help of the Eigenvalue Criterion or the Kaiser Criterion, the Eigenvalues of the individual factors are needed. If these are calculated, all factors with eigenvalues greater than 1 are used.

17.6.2 Scree-Test

In order to determine the number of factors with the help of the scree test or scree plot, the eigenvalues are sorted by size and represented by a line chart. Where there is a bend in the chart, the number of factors can be read.

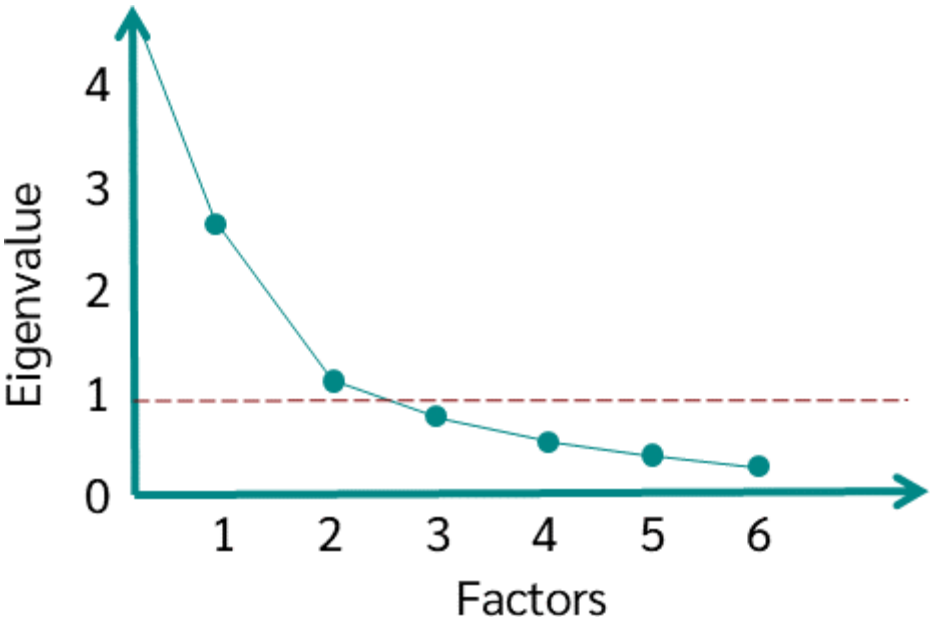


Figure 118: Scree-Test

Furthermore, in the table "Explained total variance" the variance can be read, which explains each individual factor and the cumulative variance.

Explained total variance

Copy / /

Component	Total	% of variance	Accumulated %
1	1.873	31.213	31.213
2	1.483	24.714	55.927
3	1.362	22.698	78.625
4	0.674	11.233	89.858
5	0.402	6.704	96.563
6	0.206	3.437	100

Factor 1 explains 31.21% of the total variance

Factor 2 explains 24.71% of the total variance

78.625% of the total variance can be explained by the first three factors

17.6.3 Communalities

Once the number of factors is determined, the communalities can be calculated. As written above, the communality indicates the variance of the variables, which is explained by all factors. If e.g. three factors were selected, the communalities give the variance portion of the respective variable at that with these three factors to be described can.

Communalities

Copy / /

	Extraction
outgoing	0.775
sociable	0.883
hard-working	0.667
dutiful	0.723
warm-hearted	0.859
helpful	0.811

Of the variable **outgoing 77.5%** of the variance can be described with the three factors

Of the variable **sociable, 88.3%** of the variance can be described with the three factors

...

17.6.4 Component matrix

The component matrix indicates the factor loads of the factors on the variables. Since the first factor explains most of the variance, the values of the first component or factor are the largest. With this form of representation it is however difficult to make a statement about the factors, therefore this matrix is still rotated.

Component matrix

Copy /

	Component		
	1	2	3
outgoing	0.673	0.185	-0.536
sociable	0.803	0.472	-0.123
hard-working	0.155	0.115	0.793
dutiful	0.537	0.072	0.656
warm-hearted	-0.172	0.911	0.005
helpful	-0.658	0.614	0.022

17.6.5 Rotation Matrix

The computation of the component matrix has the consequence that on the first factor many variables highly load. This results in the fact that the component matrix usually cannot be interpreted meaningfully. Therefore, a rotation of this matrix takes place. For this rotation there are different procedures, but the most common is the analytical Varimax rotation.

17.6.6 Varimax Rotation

With the help of the Varimax rotation it should be analytically ensured that per factor certain variables load as high as possible and the other variables load as low as possible. This is obtained when the variance of the factor charges per factor should be as high as possible.

Rotated Component Matrix (Varimax)

Copy /

	Component		
	1	2	3
outgoing	0.839	-0.139	-0.227
sociable	0.905	0.067	0.245
hard-working	-0.131	0.047	0.805
dutiful	0.219	-0.162	0.805
warm-hearted	0.232	0.893	0.088
helpful	-0.298	0.842	-0.114

Here it is to be recognized now that "outgoing" and "sociable" lay on Extraversion, "industriously" and "dutiful" lay on conscientiousness and "warmheartedly" and "helpfully" on agreeableness.

18. Cluster Analysis

A hierarchical cluster analysis is a clustering method that creates a hierarchical tree or dendrogram of the objects to be clustered.

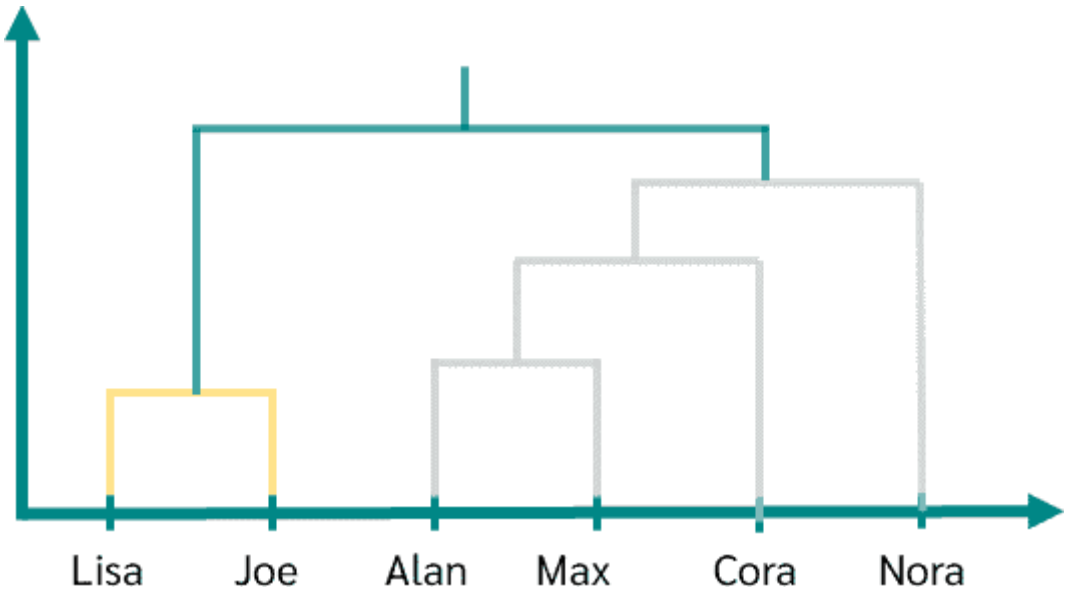


Figure 119: Cluster Analysis example

The tree represents the relationships between objects and shows how objects are clustered at different levels.

18.1 Example Hierarchical Cluster Analysis

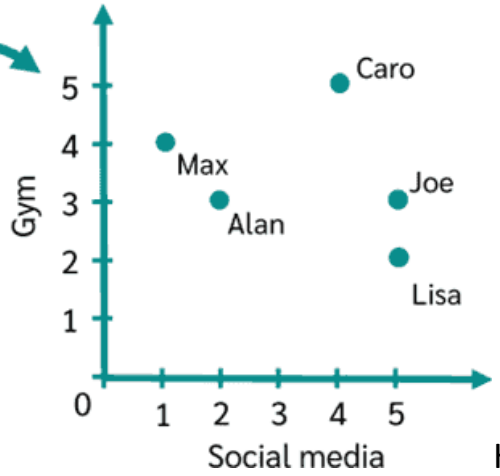
Example: We asked people about how many hours a week they spend on social media platforms and at the gym.

	Social media	Gym
Alan	2	3
Lisa	5	2
Joe	5	3
Max	1	4
Cora	4	5

We now want to know if there are clusters in this dataset and perform a Hierarchical Cluster Analysis.

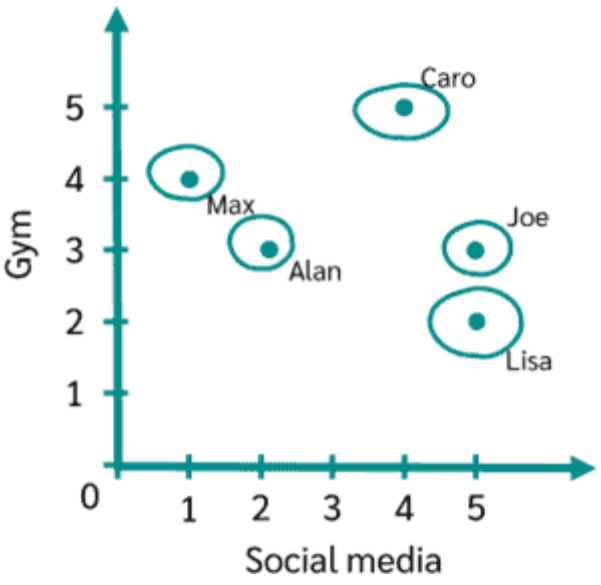
18.2 Calculating a Hierarchical Cluster Analysis

	Social media	Gym
Alan	2	3
Lisa	5	2
Joe	5	3
Max	1	4
Cora	4	5



How is a

With this we can now start to create the clusters. In the first step we assign a cluster to each point. So we have as many clusters as we have persons.



The goal now is: to merge more and more clusters little by little, until finally all points are in one cluster.

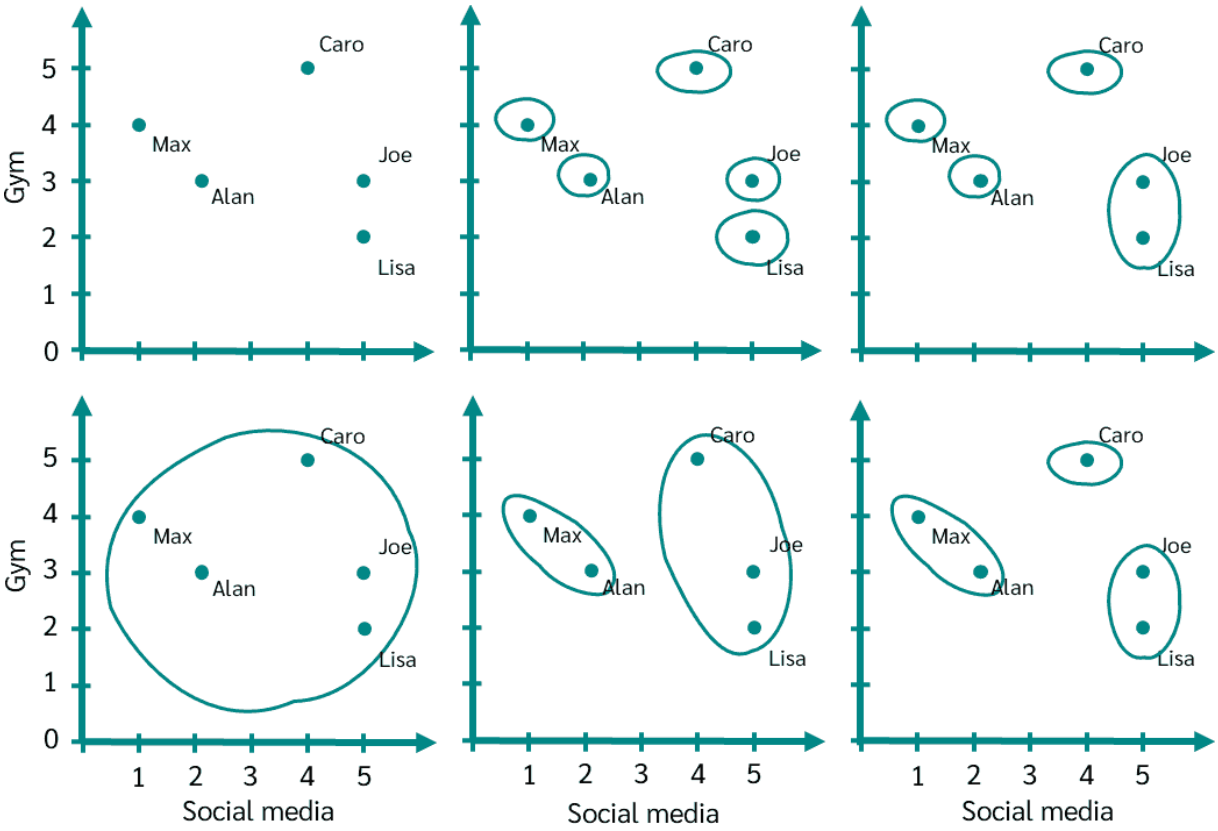


Figure 120: Cluster Analysis steps

In each step, the clusters that are closest together are always merged. What does "closest together" mean?

For this we need to determine two things:

- How the distance between two points is measured.
- How points in a cluster are connected.

18.3 Distance between two points

Let's start with the question, how do we calculate the distance between two points? Here are the most known distances:

- the Euclidean distance,
- the Manhattan distance
- and the Maximum distance.

Let's take the distance between Max and Caro. The difference on the y-axis is 1 and the difference on the x-axis is 4.

18.3.1 Euclidean Distance

The Euclidean distance is the square root of the sum of the squared differences.

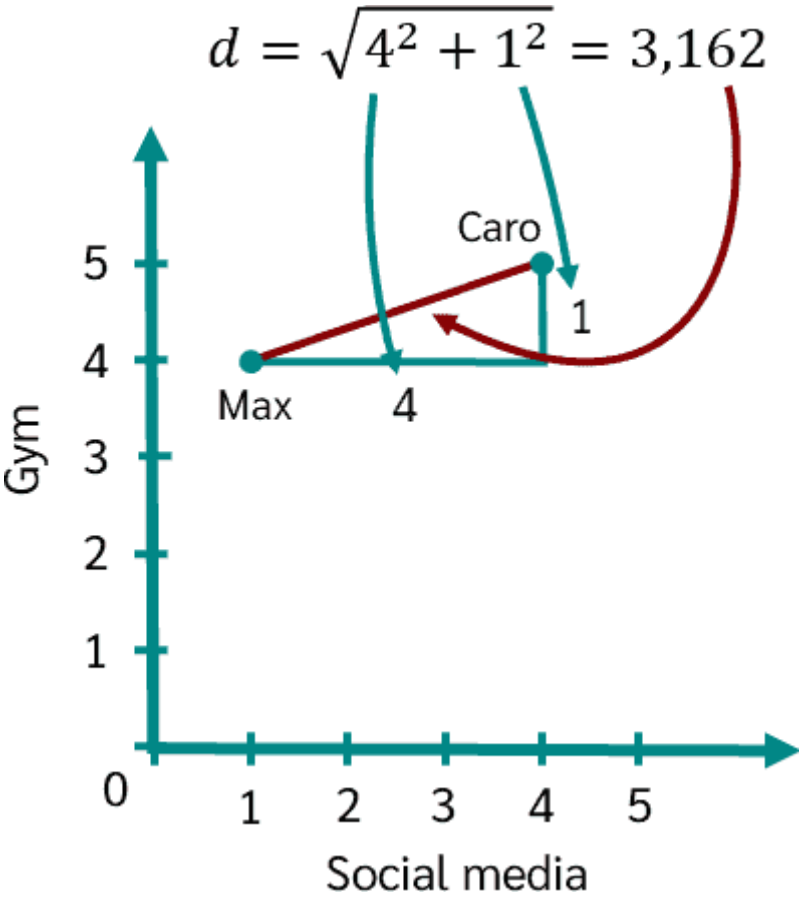
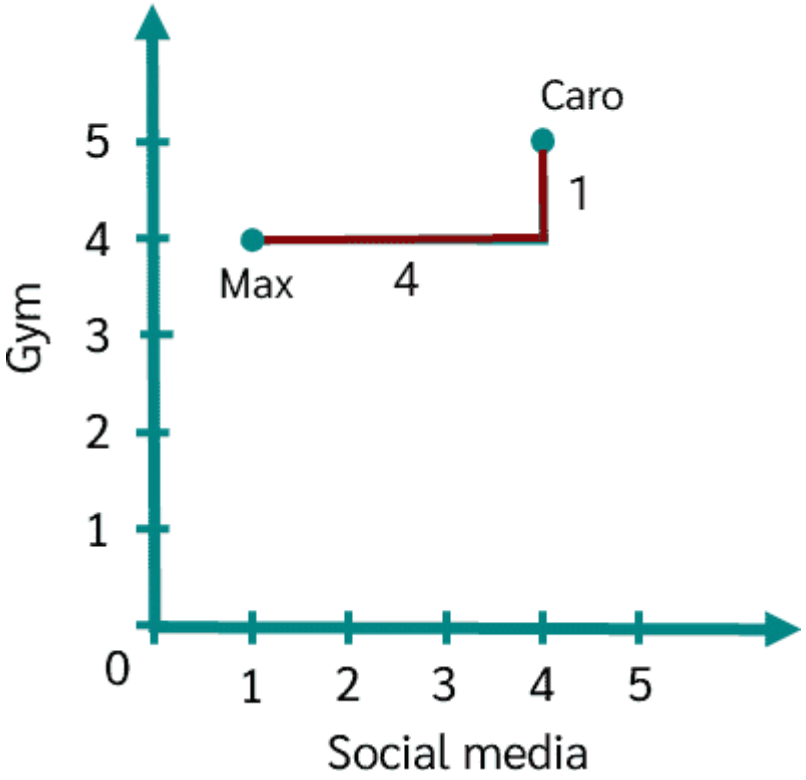


Figure 121: Euclidean Distance

18.3.2 Manhattan Distance

The Manhattan distance uses the sum of the absolute differences. So we simply calculate 4 plus 1 and keep a distance of 5.

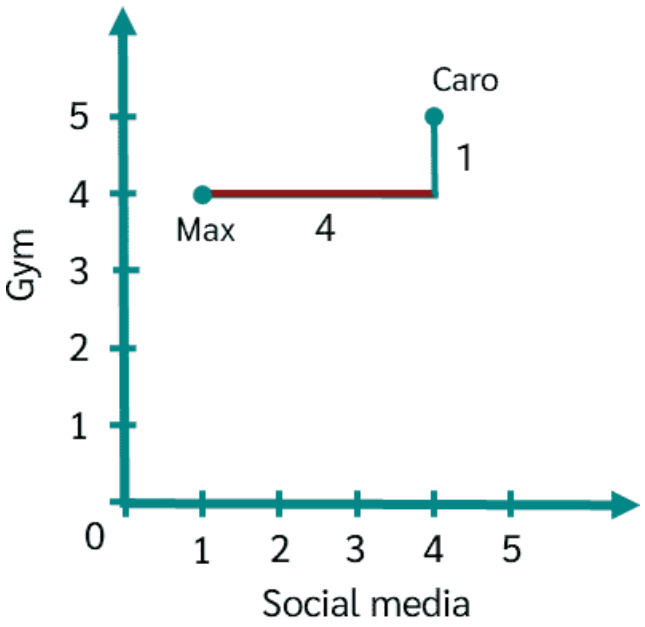
$$d = 4 + 1 = 5$$



18.3.3 Maximum Distance

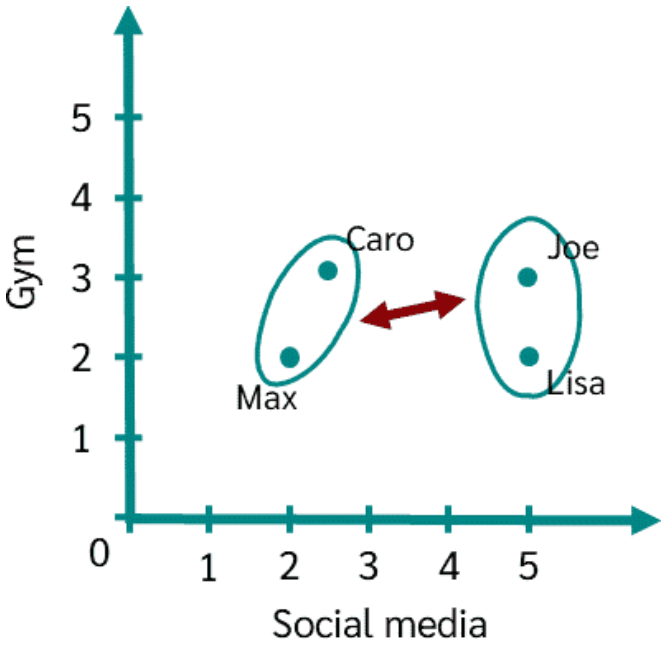
The maximum distance is simply the maximum value of the absolute differences. In this case it is 4.

$$d = max(4\,,1) = 4$$



18.4 Linking methods

Now that we know what ways there are to calculate the distances between points, we need to determine how to link the points within a cluster.

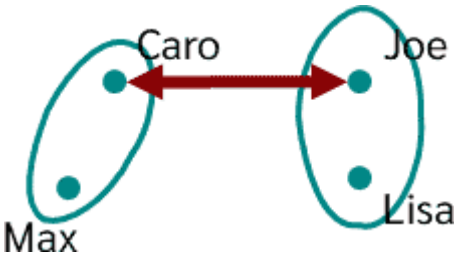


Let's say we have a cluster with the points **Joe and Lisa** and a cluster with **Max and Caro**. Now how do we determine the distance between these two clusters? Here are the most popular methods:

- Single-linkage,
- Complete-linkage
- and Average-linkage.

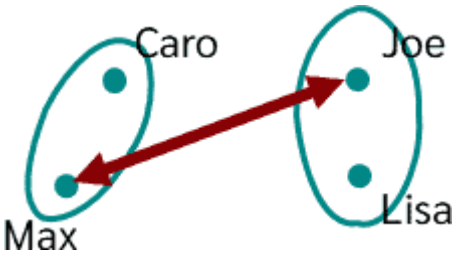
18.4.1 Single-linkage

Single-linkage uses the distance between the closest elements in the cluster. This is the distance between Caro and Joe.



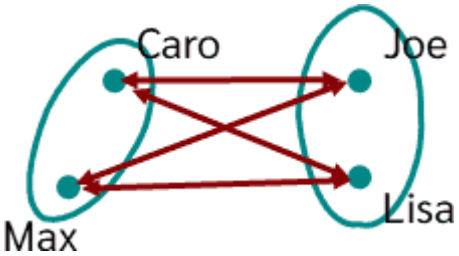
18.4.2 Complete-linkage

Complete linkage uses the distance between the farthest elements in the cluster. So between Max and Joe.



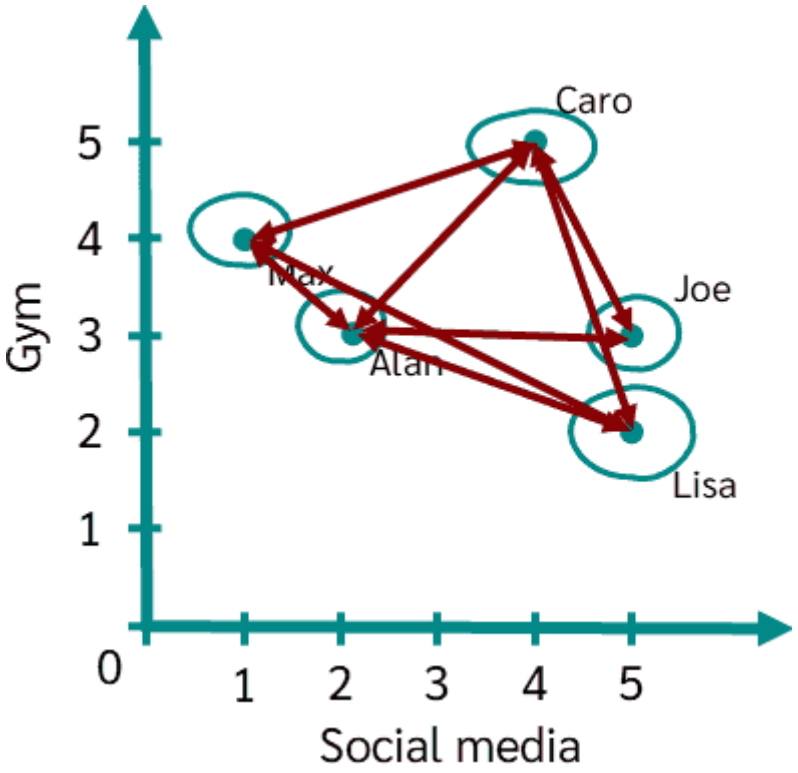
18.4.3 Average-linkage

Average-linkage uses the average of all pairwise distances. From each combination the distance is calculated and from it the average.



18.5 Example Hierarchical Cluster Analysis

For our example we use the Euclidean distance and the single-linkage method. So now we need the distance from each cluster to the other clusters.



	Alan	Lisa	Joe	Max	Caro
Alan	0				
Lisa	3.16	0			
Joe	3.00	1.00	0		
Max	1.41	4.47	4.12	0	
Caro	2.83	3.16	2.24	3.16	0

The distance between Alan and Lisa is given by:

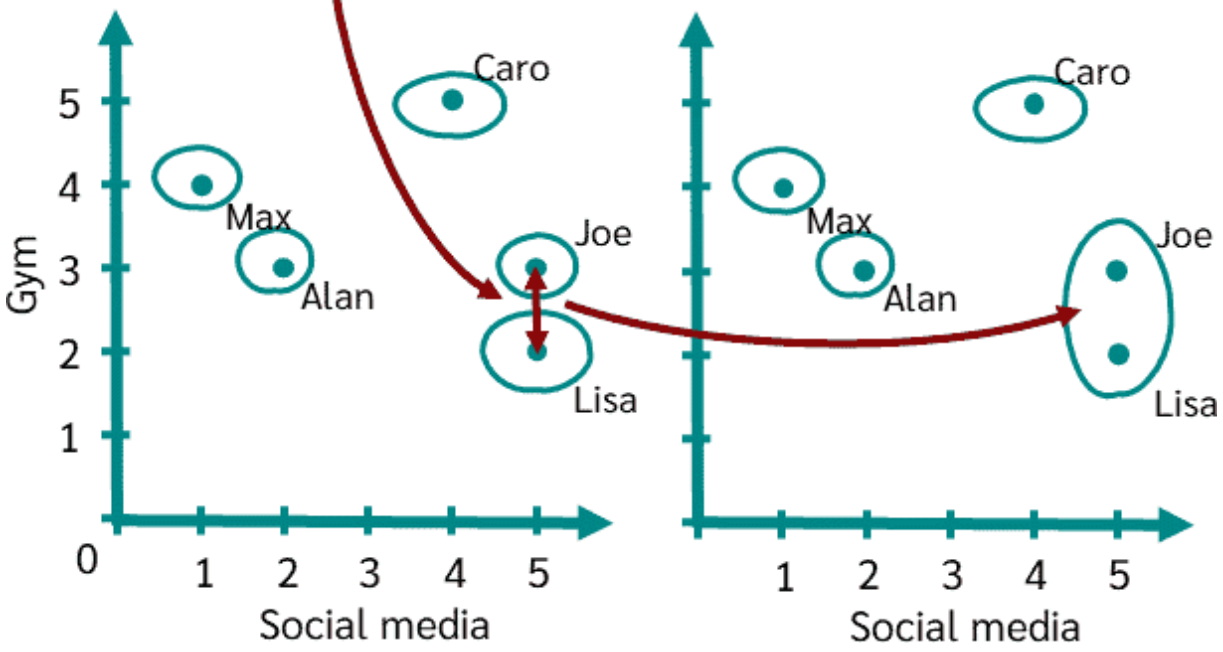
$$d = \sqrt{(5-2)^2+(2-3)^2} = 3.16$$

	Social media	Gym
Alan	2	3
Lisa	5	2
Joe	5	3
Max	1	4
Caro	4	5

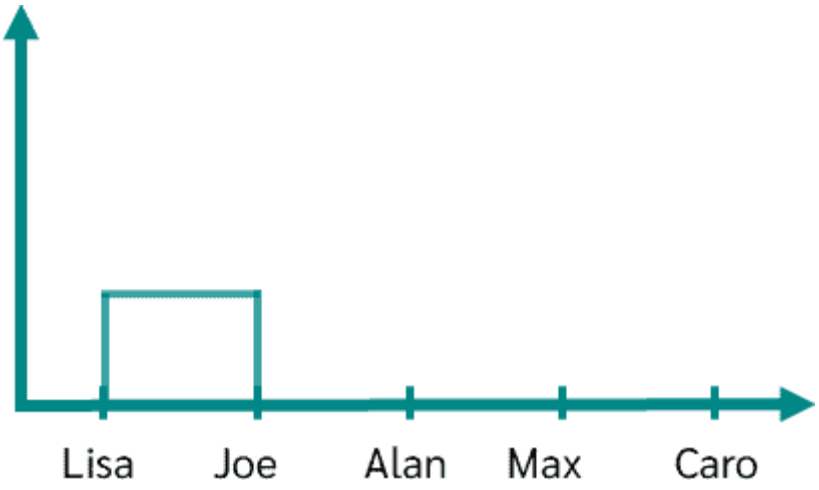
	Alan	Lisa	Joe	Max	Caro
Alan	0				
Lisa	3.16	0			
Joe	3.00	1.00	0		
Max	1.41	4.47	4.12	0	
Caro	2.83	3.16	2.24	3.16	0

We can now do this for all other combinations until we have calculated the total distance matrix. Now we can merge the first clusters. For this we look between which two clusters we have the smallest distance. This is the case between Joe and Lisa.

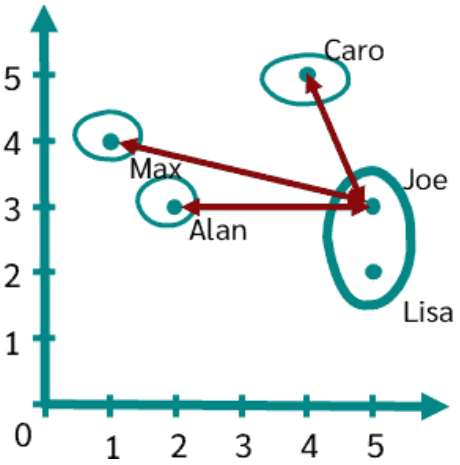
	Alan	Lisa	Joe	Max	Caro
Alan	0				
Lisa	3.16	0			
Joe	3.00	1.00	0		
Max	1.41	4.47	4.12	0	
Caro	2.83	3.16	2.24	3.16	0



With this, we now combine Joe and Lisa into one cluster. In our tree diagram or dendrogram we can draw the first connection.



Now we need to update our distance matrix. We decided to use the single linkage method. So the distance between two clusters is given by the elements that are closest to each other. To the clusters Alan, Max and Caro, from the cluster Lisa and Joe respectively, Joe is always the closest person.

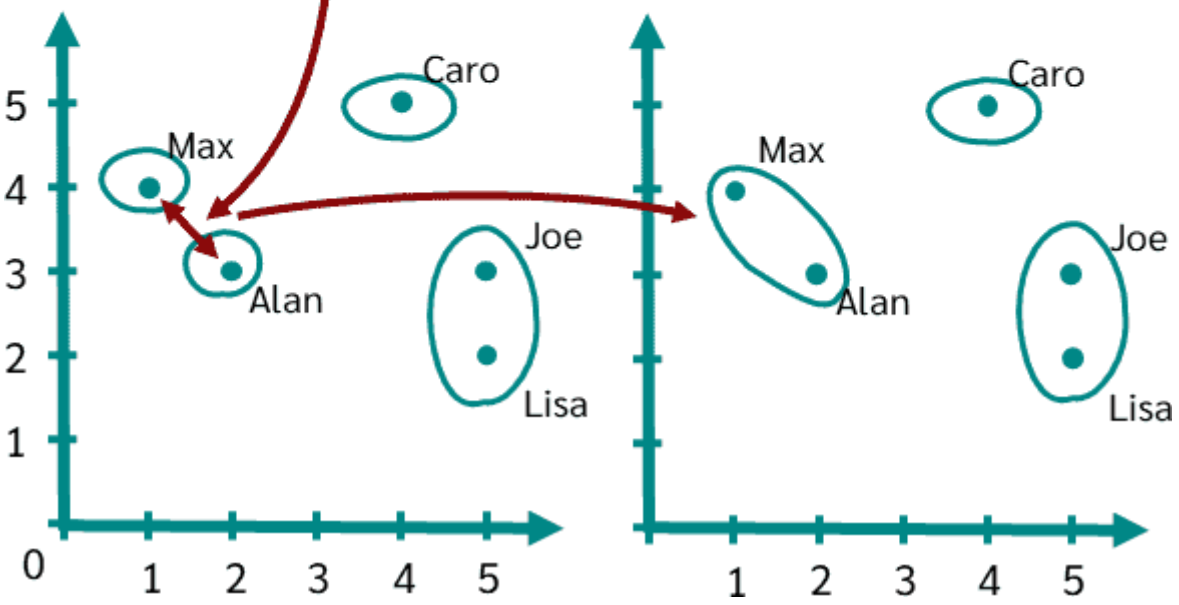


	Alan	Lisa, Joe	Max	Caro
Alan	0			
Lisa, Joe	3.00	0		
Max	1.41	4.12	0	
Cora	2.83	2.24	3.16	0

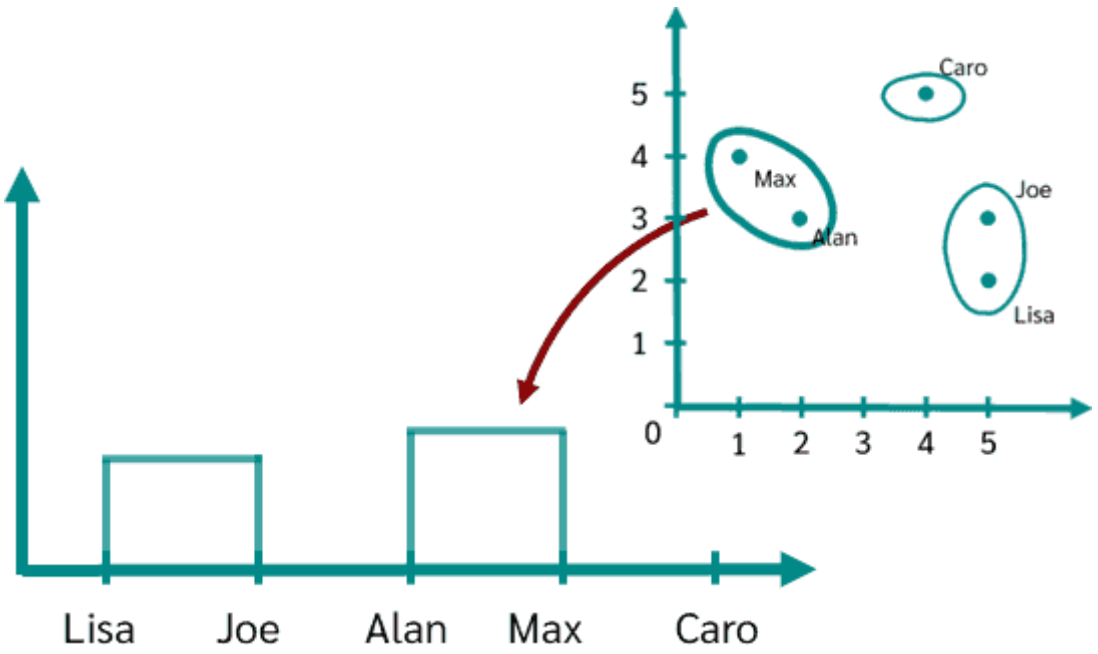
So we calculate the distance from Alan to Joe, the distance from Max to Joe, and the distance from Caro to Joe.

Now we again merge the clusters that are closest. These are Max and Alan.

	Alan	Lisa, Joe	Max	Caro
Alan	0			
Lisa, Joe	3.00	0		
Max	1.41	4.12	0	
Caro	2.83	2.24	3.16	0

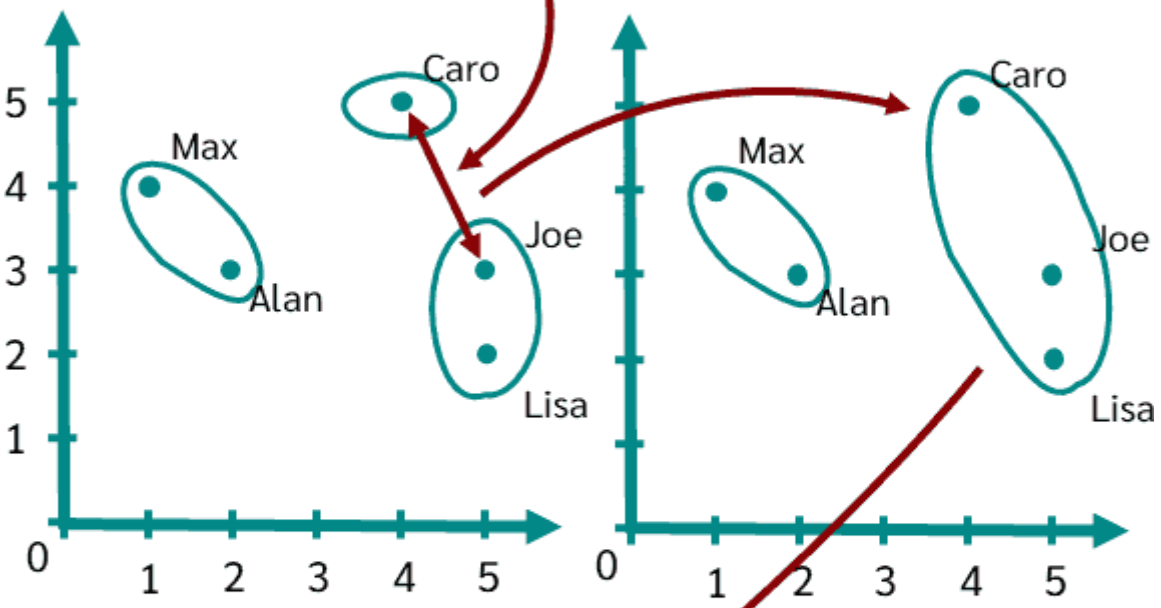


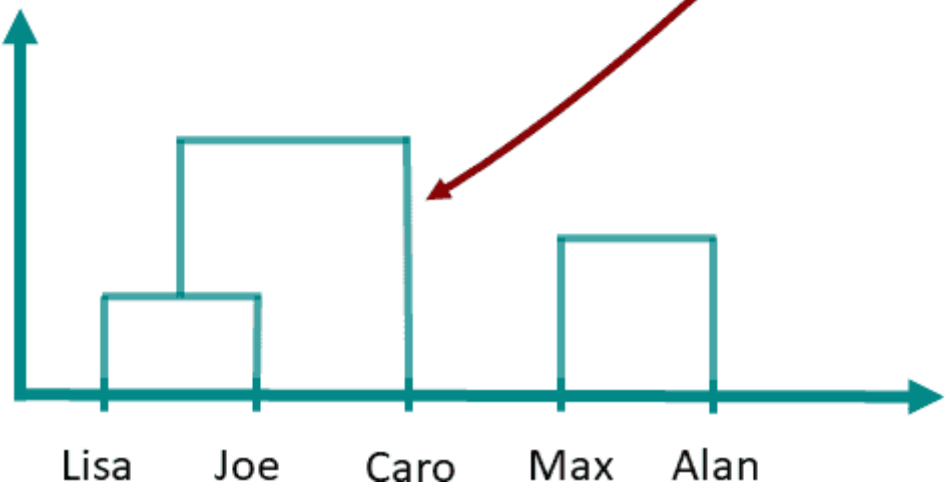
In our tree diagram or dendrogram, we can draw in the second connection.



Now we update the distance matrix again. We calculate the distance between Alan and Joe, Caro and Joe and between Caro and Alan. We get the smallest distance between the Caro cluster and the Lisa and Joe cluster.

	Lisa, Joe	Max, Alan	Caro
Lisa, Joe	0		
Max, Alan	3,00	0	
Caro	2,24	2,83	0

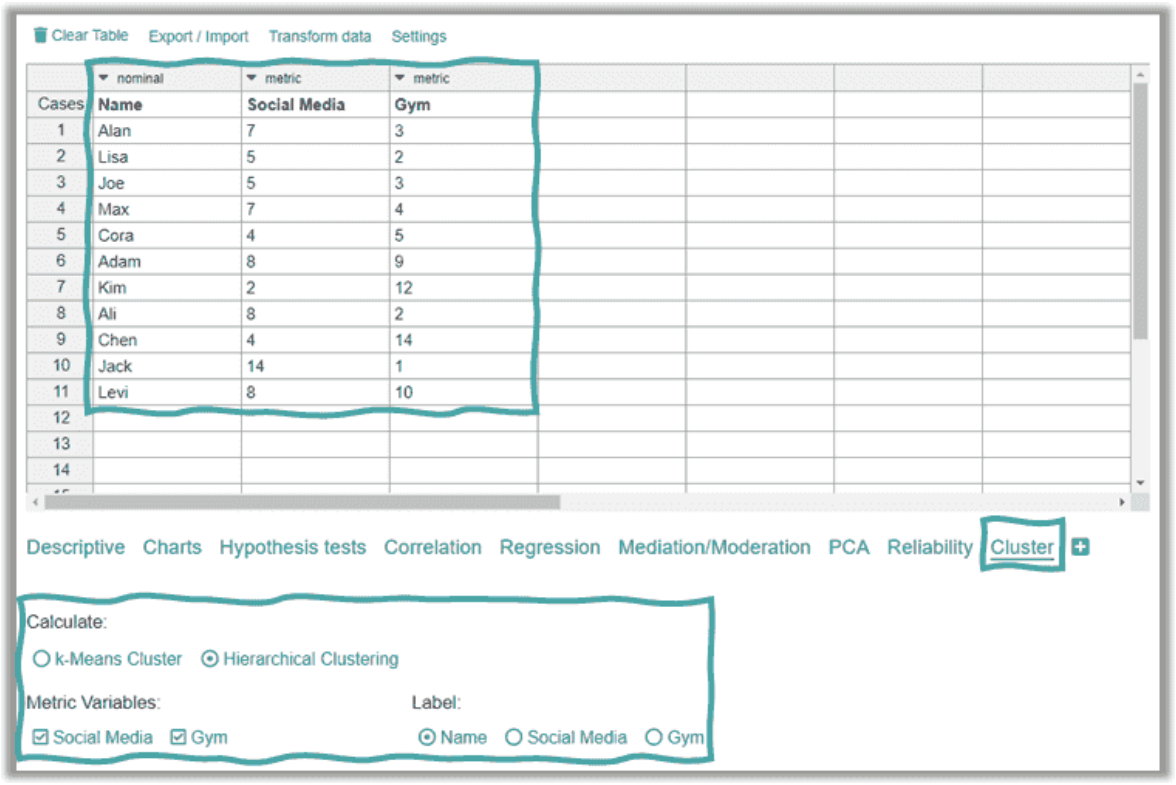




18.5.1 Calculate hierarchical cluster analysis with DATAtab

To calculate a hierarchical cluster analysis online, just visit the statistics calculator and copy your own data into the table or use the link to load the dataset. Now we click on cluster and select hierarchical cluster.

If we now click on Social Media and Gym a hierarchical cluster analysis will be calculated for us. Additionally we can specify the label, in our case the names of the persons.



Now we can specify which connection method should be used and how the distance should be calculated. We simply take Single linkage and the Euclidean distance again.

Hierarchical cluster analysis

Linkage method:

(•) Single-linkage ( ) Complete-linkage ( ) Average-linkage

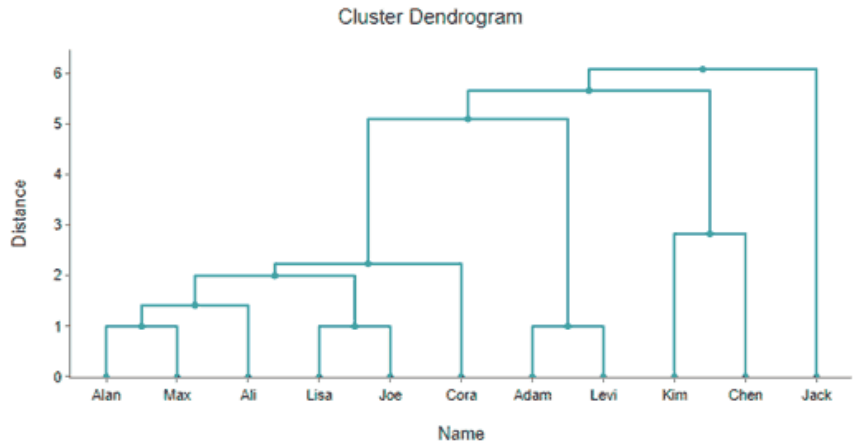
Distance:

(•) Euclidian ( ) Manhattan ( ) Maximum

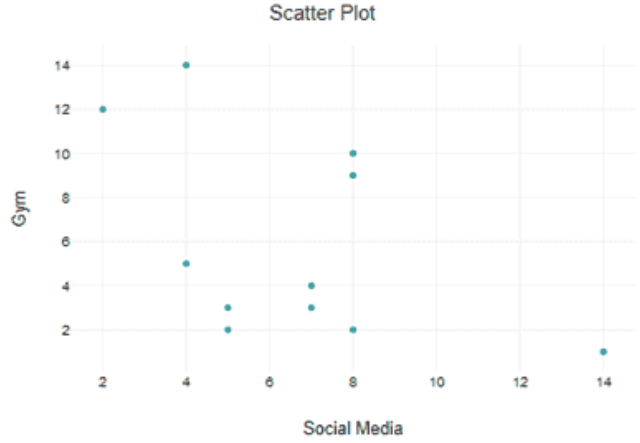
Number of Clusters:

1

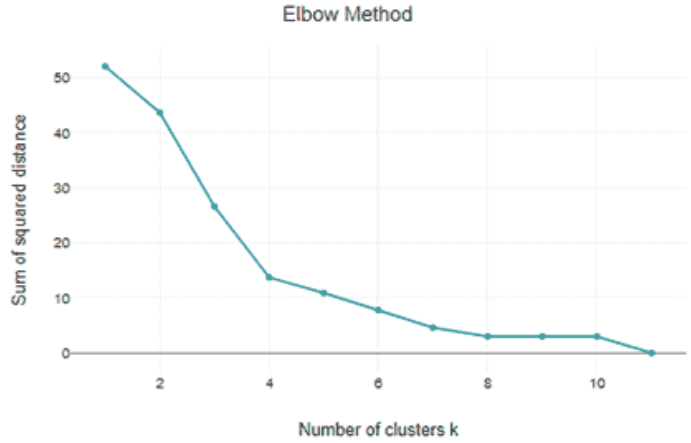
Download png Download svg Settings



Download png Download svg Settings



Download png Download svg Settings

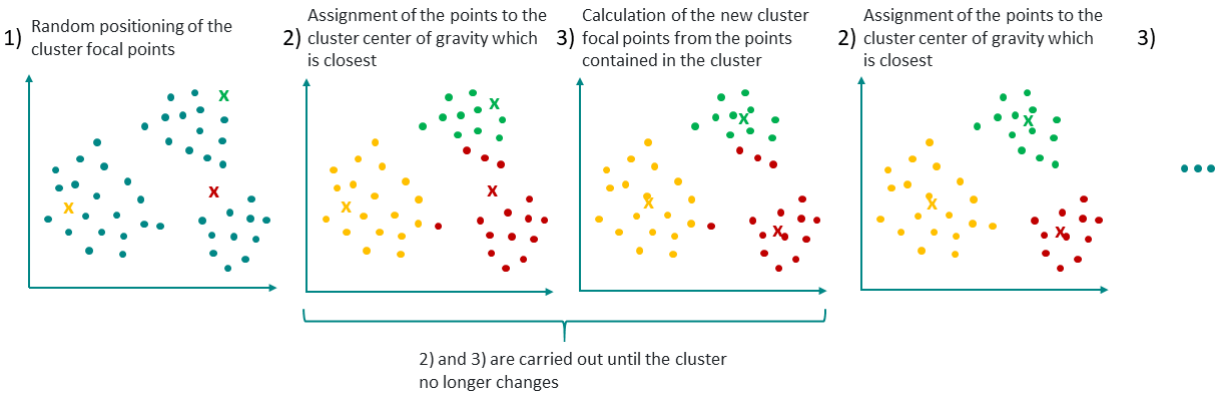


Now we get the results output down here. We see the tree plot, a scatter plot and the elbow plot. In the elbow plot we can now read how many clusters we take. We can see a kink here, so we'll take 4 as the cluster count. We can still select these up here and then in the tree plot we get the 4 clusters highlighted by different colors. We see the first cluster, the second cluster, the third cluster and the fourth cluster.

18.6 K-means cluster analysis

The k-Means method, which was developed by MacQueen (1967), is one of the most widely used non-hierarchical methods. It is a partitioning method, which is particularly suitable for large amounts of data.

- First, an initial partition with k clusters (given number of clusters) is created.
- Then, starting with the first object in the first cluster, Euclidean distances of all objects to all cluster foci are calculated.
- If an object is detected whose distance to the center of gravity of the own cluster is greater than the distance to the center of gravity (centroid) of another cluster, this object is shifted to the other cluster.
- Finally, the centroids of the two changed clusters are calculated again, since the compositions have changed here.
- These steps are repeated until each object is located in a cluster with the smallest distance to its centroid (center of the cluster) (optimal solution).

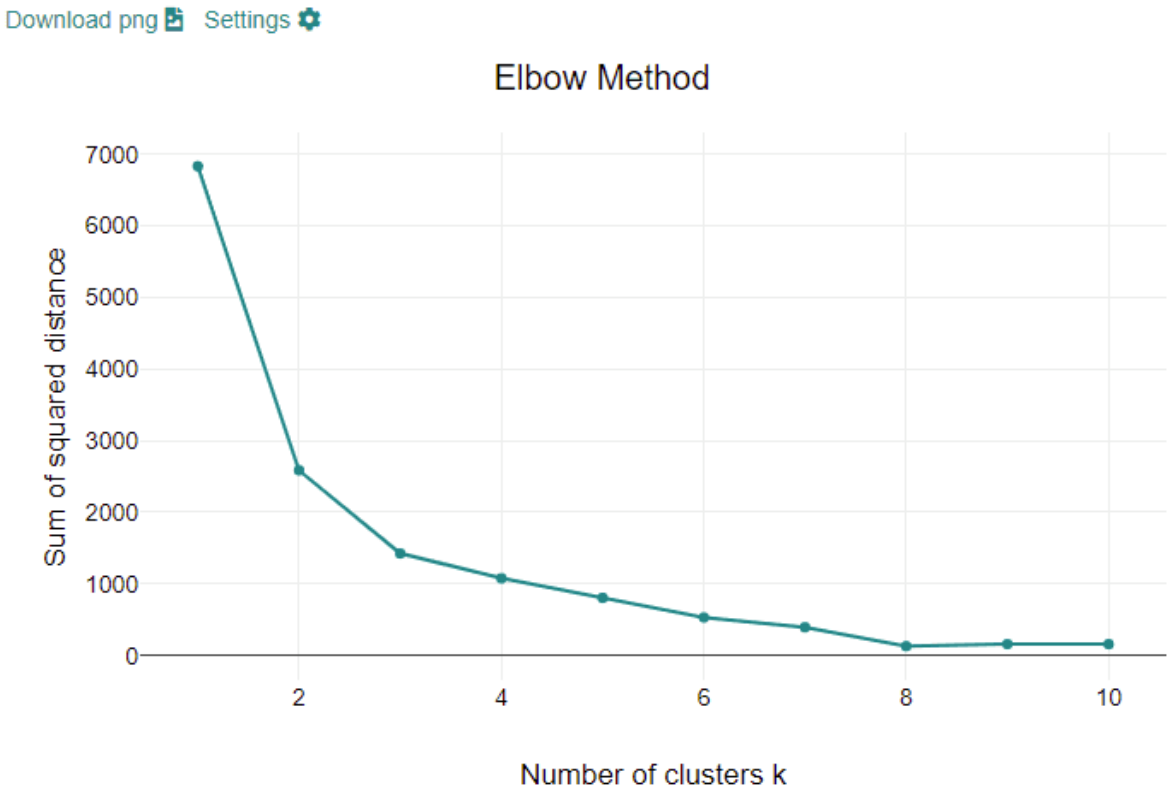

1) Random positioning of the cluster focal points
2) Assignment of the points to the cluster center of gravity which is closest
3) Calculation of the new cluster focal points from the points contained in the cluster
2) Assignment of the points to the cluster center of gravity which is closest
3)
2) and 3) are carried out until the cluster no longer changes

18.6.1 Optimal cluster number

The number of clusters in the k-Means method must be determined before the start and is therefore not determined by the cluster method. But what is the optimal number of clusters in the k-Means method? The elbow method is a common way to determine the appropriate number of clusters.

18.6.2 Elbow curve

When you want to calculate a cluster analysis, often the big question is how many clusters should I take, The Elbow Method helps with this question! With each new cluster, the total variation in each cluster becomes smaller and smaller. In the extreme case, when there are as many clusters as there are points, the result is zero. However, in most cases, the reduction of the total variation becomes smaller after a certain point. This point is then used as the optimal cluster number.



18.6.3 Scaling data for k-means clustering

If the variables under consideration do not have the same unit, it is often advisable to scale the data before cluster analysis.

18.6.4 K-means clustering calculator

Why use a K-Means Clustering Calculator? Imagine having a large dataset with thousands of data points and you need to segment this data efficiently. Instead of wrestling with complex code and algorithms, the K-Means Clustering Calculator offers a hassle-free solution:

- **User-Friendly Interface:** No coding expertise? No problem! Navigate and use with ease.
- **Accurate Results:** Harness the power of advanced algorithms to get precise cluster assignments.
- **Timesaving:** Why spend hours segmenting data manually when you can do it in minutes?

18.6.5 Key Features

- **Data Visualization:** Get a clear view of how your data is clustered with interactive graphs.
- **Elbow Method Integration:** Not sure about the optimal number of clusters? The calculator employs the Elbow Method to suggest the best 'k' for your data.
- **Downloadable Outputs:** Extract your results in various formats for further analysis.

19. Market Basket Analysis [Association Analysis]

Market basket analysis (also called association analysis) is one of the most important methods used to uncover relationships between items. It looks for combinations of items that frequently occur together in transactions. In other words, it enables retailers to identify relationships between the items that customers buy.

19.1 What does association analysis do?

Let's say you have set up your own online clothing store. Your goal now is to achieve the highest possible turnover with this store.

Figure 122: Market basket analysis 1

In order to achieve the highest possible sales, you naturally want every customer to buy as much as possible. One way to motivate the customer to buy more products is to suggest more products. The big question now is: Which product do I best suggest to the customer? This is where market basket analysis or association analysis comes into play.

Figure 123: Market basket analysis 2

The market basket analysis gives an answer to the question: How likely is it that a customer buys product A if he already has product B in his market basket.

The market basket analysis tells you which products or goods are often bought together. So if a customer already has a pair of pants and shoes in the shopping cart, how likely is it that this customer will also buy a shirt, socks or a t-shirt.

19.2 Market Basket Analysis Example

To calculate a shopping cart analysis, you need a list of past purchases, where you can see which products were bought together in one purchase.

Figure 124: Market basket analysis 3

So you have the respective products listed and each row is a transaction. Let's say that is your example data, you have the products jeans, shirt, jacket and shoes.

Jeans	Shirt	Jacket	Shoes	Transactions
1	1	0	1	Jeans, Shirt ,Schuhe
0	0	0	1	Schuhe
0	1	0	1	Shirt, Schuhe
...	...	...	...	...
1	1	0	1	Jeans ,Shirt ,Schuhe
0	1	0	1	Shirt ,Schuhe

Each row is a transaction or a purchase. 1 means bought, 0 means not bought. So the first person bought jeans, shirt and shoes.

Now, so that we have results that we can interpret, let's first calculate a market basket analysis using DATAtab for this data. To do this, go to the market basket analysis calculator on DATAtab and copy your data into the table.

Now we can specify a minimum support and a minimum confidence. For this data DATAtab issued us these association rules:

Rules

Copy Word Copy Excel

Lhs	Rhs	Frequency	Support	Confidence	Lift
Shirt	Shoes	8	0.42	0.8	1.27
Jeans, Shirt	Shoes	3	0.16	0.75	1.19
Shoes	Shirt	8	0.42	0.67	1.27
Jeans, Shoes	Shirt	3	0.16	0.6	1.14

The association rules are in the form: If the products under Lhs (Left hand side) are present in a transaction, then the products under Rhs (Right hand side) are also present with some probability.

19.3 Interpreting the results of a Market basket analysis

We look at the results of the basket analysis using the first set of association rules.

Jeans	Shirt	Jacket	Shoes
1	**1**	0	**1**
0	0	0	**1**
0	**1**	0	**1**
0	**1**	1	**1**
1	0	1	0
1	0	1	0
1	0	1	0
1	0	0	0
0	**1**	0	**1**
0	**1**	0	**1**
0	0	1	**1**
0	0	1	0
0	**1**	1	0
1	**1**	1	0
1	0	0	**1**
1	0	0	**1**
1	**1**	0	**1**
1	**1**	0	**1**
0	**1**	0	**1**

Lhs	Rhs	Frequency	Support	Confidence	Lift
Shirt	Shoes	8	0.42	0.8	1.27

Regel

Lhs → *Rhs*
Shirt → Shoes

Frequency

frq (Lhs , Rhs)
frq (Shirt , Shoes) = 8

Support

frq(Lhs , Rhs) / N
8 / 19 = 0.42

Confidence

frq (Lhs , Rhs) / frq (Lhs)
8 / 10 = 0.8

Lift

$$\frac{Support}{Support(\ Lhs\ )\ x\ Support(\ Rhs\ )}$$

$$\frac{Support}{Support(\ Shirt\ )\ x\ Support(\ Shoes\ )}$$

$$\frac{8\ /\ 19}{10\ /\ 19\ x\ \ 12\ /\ 19} = 1.27$$

19.3.1 Frequency

The frequency in the results table tells us how often the products under Lhs and Rhs occur in a transaction, so in our case, how often does shirt and shoes occur in a transaction.

So let's just count through how many transactions both occur in, which is 8 transactions.

19.3.2 Support

Support tells us what percentage of all transactions that is, or in other words, how likely it is that shirt and shoes will occur in a transaction. So we just divide the frequency by the number of all transactions.

19 transactions we have in total, so we get 8/19, which is equal to 0.42. So the probability of shirt and shoes occurring in a transaction is 42 percent.

19.3.3 Confidence

Confidence now tells us, if the products under Lhs are in an order, how likely it is that the products under Rhs are then also in the shopping cart.

In our example this means: How likely is it that if shirt occurs in the cart, then shoes are also in the cart. We can calculate this by dividing the frequency of shirt and shoes by the frequency of shirt.

19.3.4 Lift

And finally, the lift. The lift indicates the factor by which the probability of buying the products under Rhs increases if the products under Lhs have already been bought. So in our example. If the product Shirt is in the shopping cart, it is 1.27 times more likely that Shoes will be purchased than if the product Shirt is not in the shopping cart.

19.3.5 Market basket analysis and data mining

Shopping cart analysis is a method from the field of data mining. Depending on how much data is available, the analysis can be very computationally intensive.

However, with The Apriori Algorithm, there are very effective methods to efficiently determine the association rules.

19.3.6 Critical note on the market basket analysis

Let's say your market basket analysis shows that if a person buys a pair of pants and shoes, there is a high probability that they will also buy a shirt. Now you suggest a shirt to all customers who buy pants and shoes. This increases the probability that a shirt will be bought under this condition and another market basket analysis will be falsified.

20. Cronbach’s Alpha

Cronbach's Alpha (or tau-equivalent reliability) is a measure of the relationship between a group of questions. The group of questions is called a scale and each question in the group is an item. Cronbach's alpha is therefore a measure of the internal consistency of a scale and therefore of the strength of its reliability.

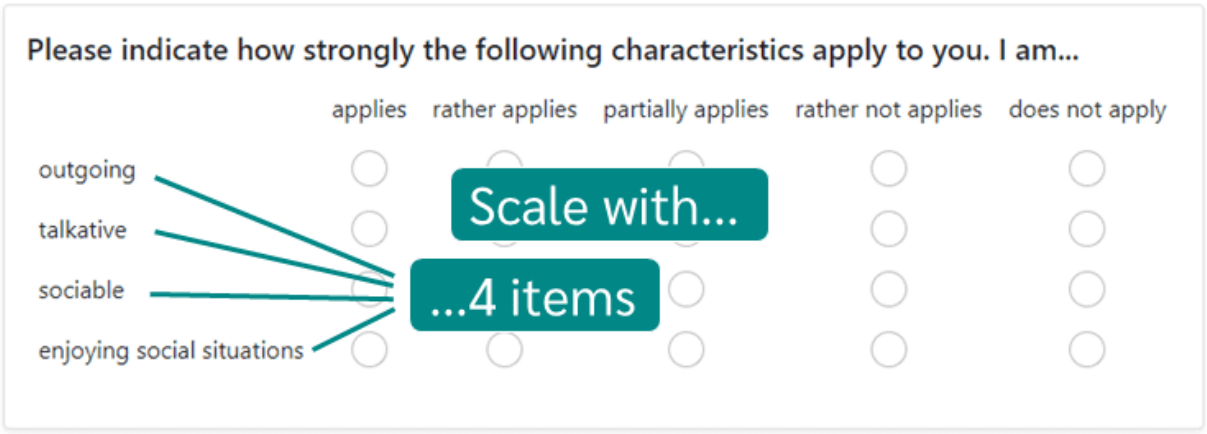


Figure 125: Scale with 4 items

Cronbach's Alpha is the correlation between the answers in a questionnaire and can take values between 0 and 1. The higher the average correlation between items, the greater the internal consistency of a test.

20.1 Latent variables

Hypotheses often contain variables that cannot be measured directly. Variables that are not directly measurable are called latent variables and are, for example, writing ability, intelligence, or the attitude toward electric cars.

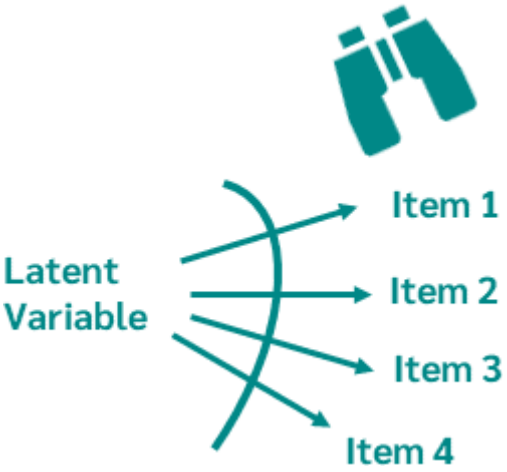


Figure 126: Latent variable

In order to make latent variables "measurable", a scale is used. A scale is a group of questions used to collectively measure a latent variable.

The goal now is that the answers to the different items match well, i.e. correlate highly. Each individual question should correlate as highly as possible with every other question.

Reliability indicates how reliably or accurately a questionnaire or test measures a true value. Reliability therefore means how accurately a test can measure a variable. The less measurement error there is, the more reliable a test is.

Cronbach's Alpha is therefore a measure of the extent to which the group of questions are related to each other and thus provides an estimate of how good or poor the measurement accuracy, known as reliability, of a group of items is.

20.2 Assumptions for Cronbach's Alpha

In the context of classical test theory, the focus is on the measurement errors that exist when a value is measured. In order to calculate Cronbach's Alpha, two conditions must be met.

- The error proportions of the items must be uncorrelated, i.e. the error proportion of one item must not be influenced by the error proportion of another item.
- The items must have the same proportion of true variance.

In practice, however, neither of these conditions is usually met. Furthermore, the more items a scale has, the higher the alpha value will be.

It is important to note that Cronbach's Alpha does not test whether each item is actually influenced by only one or more latent variables! A high value of is not evidence that the items are influenced by only one latent variable.

For the reliability of the scale to be estimated using Cronbach's Alpha, the condition that all questions or items measure the same latent variable must be met!

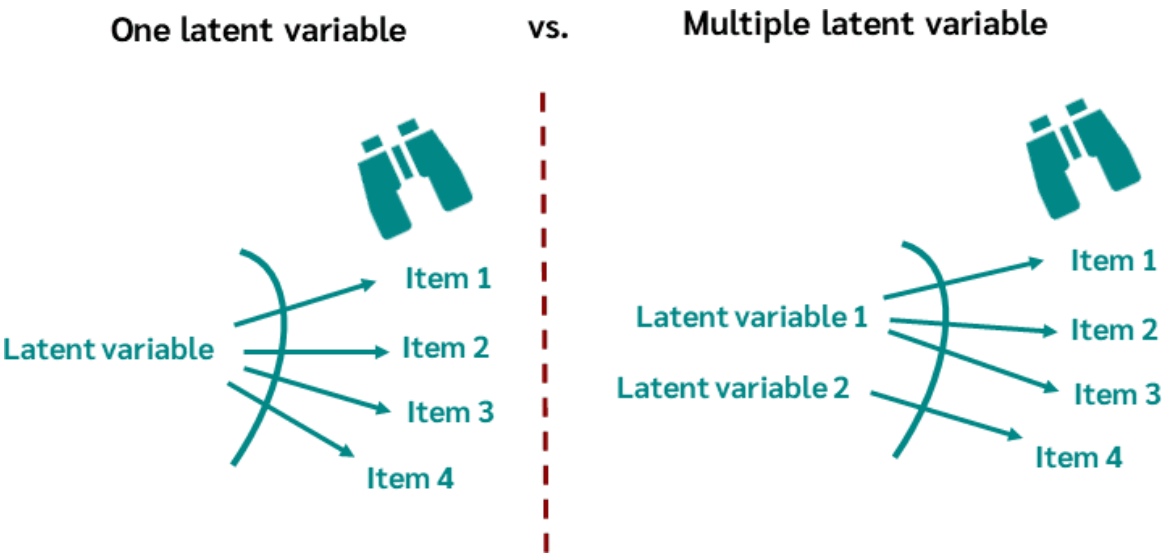
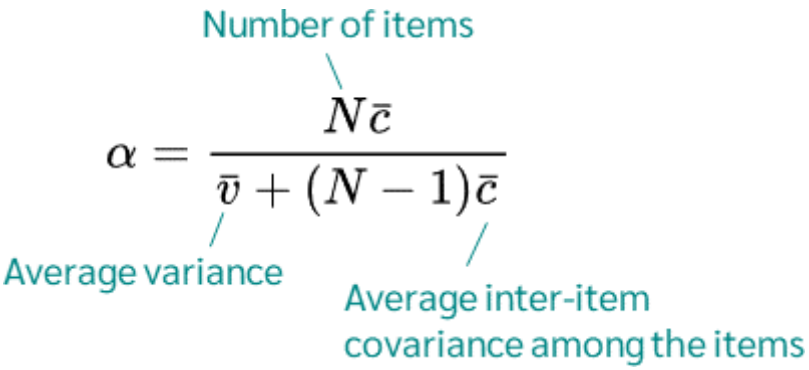


Figure 127: One and multiple latent variable(s)

In other words, if all the items measure the same latent variable, then Cronbach's Alpha tells us how well these items measure the latent variable.

20.3 Calculate Cronbach's Alpha

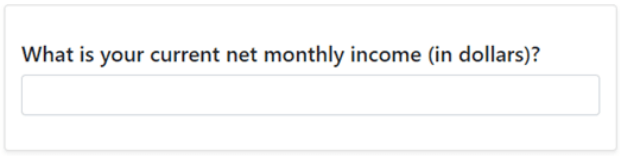
Cronbach's Alpha can be calculated using the following formula:



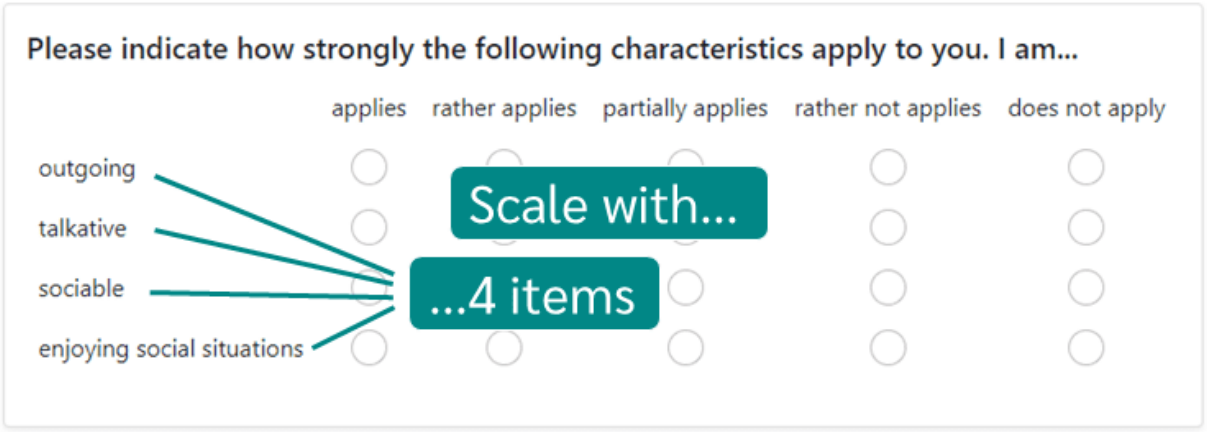
Cronbach's Alpha therefore increases as the number of items increases and as the inter-item correlation increases. correlation between the items increases. Cronbach's Alpha becomes smaller when the average inter-item correlation becomes smaller.

20.4 Example Cronbach's Alpha

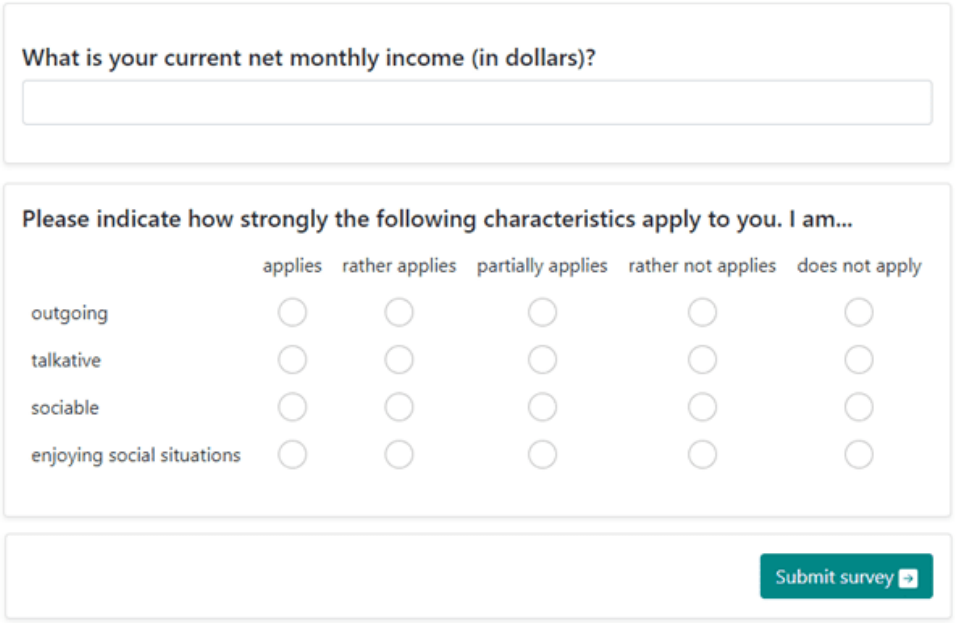
Let's say your hypothesis is: Extroverts earn more than introverts. How do you measure salary? That is easy! Just ask in the questionnaire!



But how is extraversion measured in people? Through a literature research you have discovered that Extraversion can be measured by the following scale from the Big Five Personality Traits.



So, you create a survey on datatab.net, send it out and get the answers in an Excel spreadsheet.



The sample dataset can be downloaded from datatab.net.

The four variables can now be combined into a construct that gives you a value for your unmeasurable latent variable. For example, you could do this with a sum index or a mean index.

Before that, of course, we need to check to what extent these items represent the same thing, i.e., how high Cronbach's Alpha is and how reliable the scale is.

This is done by copying the data into the upper table of the Cronbach's Alpha calculator. Then the four items are selected and DATAtab calculates the reliability statistics.

Reliability Statistics

Copy Word Copy Excel

Cronbach's Alpha	Number of Items
0.71	4

For the present data a Cronbach's Alpha of 0.71 was obtained. The table of item scale statistics is then displayed. In the table you can see how the Cronbach's Alpha changes when the respective variable or item is omitted.

Item-Total Statistics

Copy Word Copy Excel

	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
outgoing	0.47	0.66
talkative	0.76	0.48
sociable	0.58	0.59
enjoying social situations	0.21	0.79

It can be seen that when item 1 is removed, the Cronbach's alpha drops to 0.66 and when item 2 is removed, the Cronbach's alpha even drops to 0.48. However, when item 4 is removed, the Cronbach's alpha increases to 0.79. Therefore, in this case it could be considered to remove item 4.

20.5 Interpret Cronbach’s Alpha

Cronbach's Alpha should not be less than 0.6. Values above 0.7 are considered acceptable. However, the Cronbach's Alpha should preferably not be much higher than 0.9, as this would mean that the questions are "too similar" and therefore you get the same answers to the questions, in which case you could omit questions that are too highly correlated, and you would not have any loss of information. The table below can be used to interpret Cronbach's Alpha.

Cronbach's Alpha	Interpretation
> 0,9	Excellent
> 0,8	Good
> 0,7	Acceptable
> 0,6	Questionable
> 0,5	Poor
< 0,5	Unacceptable

As mentioned above, internal consistency only says something about the correlation of the items, but not about whether the items fit together in terms of content. Cronbach's Alpha only checks whether the items are correlated. The researcher must therefore ensure that only items that measure the same content are used.

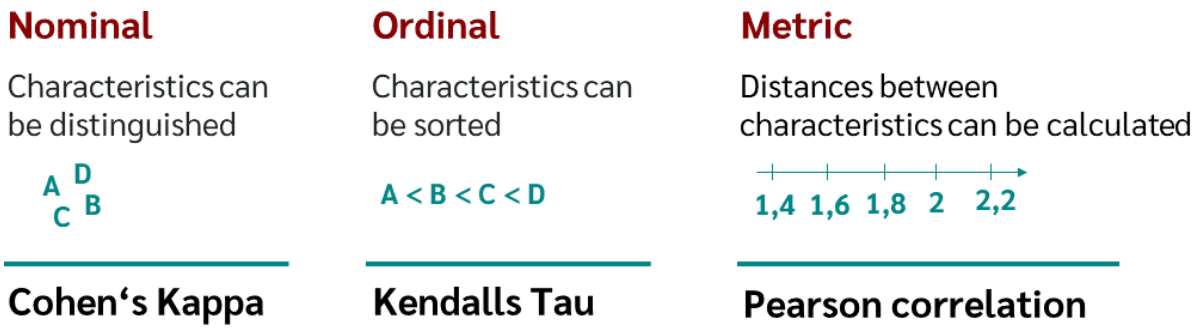
Cronbach's Alpha increases with the number of items. For example, if the scale is constructed with 8 items rather than 4, then the same correlation for the 8 items will tend to result in a larger alpha.

Furthermore, it is also important to ensure that the questions are all formulated in either a positive or a negative way. That is, a high or low value must always mean the same thing.

21. Cohen‘s Kappa

Cohen's Kappa is a measure of agreement between two dependent categorical samples, and you use it whenever you want to know if two raters' measurements are in agreement.

In the case of Cohen's Kappa, the variable to be measured by the two rates is a nominal variable.



So if you have a nominal variable and you want to know how much agreement there is between two raters, you would use Cohen's Kappa. If you have an ordinal variable and two raters, you would use Kendall's tau or the weighted Cohens Kappa, and if you have a metric variable, you would use Pearson's correlation. If you have more than two nominal dependent samples, the Fleiss Kappa is used.

21.1 Cohen’s Kappa Example

Let's say you have developed a measurement tool, for example a questionnaire, that doctors can use to determine whether a person is depressed or not. Now you give this tool to a doctor and ask her to assess 50 people with it.

For example, your method shows that the first person is depressed, the second person is depressed, and the third person is not depressed. The big question now is: Will a second doctor come to the same conclusion?

Figure 128: Cohen’s Kappa example

So, with a second doctor, the result could now look like this: For the first person, both doctors come to the same result, but for the second person, the result differs. You're interested in how big the agreement of the doctors are, and this is where the Cohen's Kappa comes in.

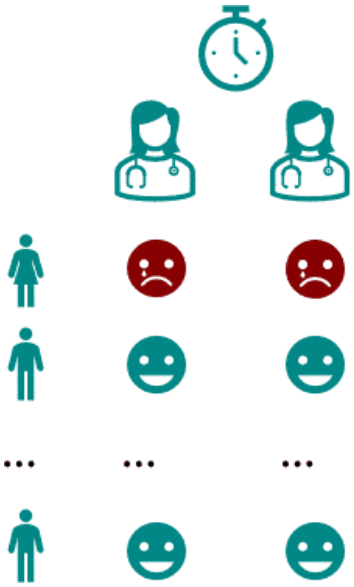
21.2 Inter-rater reliability

If the assessments of the two doctors agree very well, the inter-rater reliability is high. And it is this inter-rater reliability that is measured by Cohen's Kappa.

Definition: Cohen's Kappa is a measure of inter-rater reliability. Cohen's Kappa is therefore a measure of how reliably two raters measure the same thing.

21.3 Use cases for Cohen’s Kappa

So far, we have considered the case where two people measure the same thing. However, Cohen's Kappa can also be used when the same rater makes the measurement at two different times.

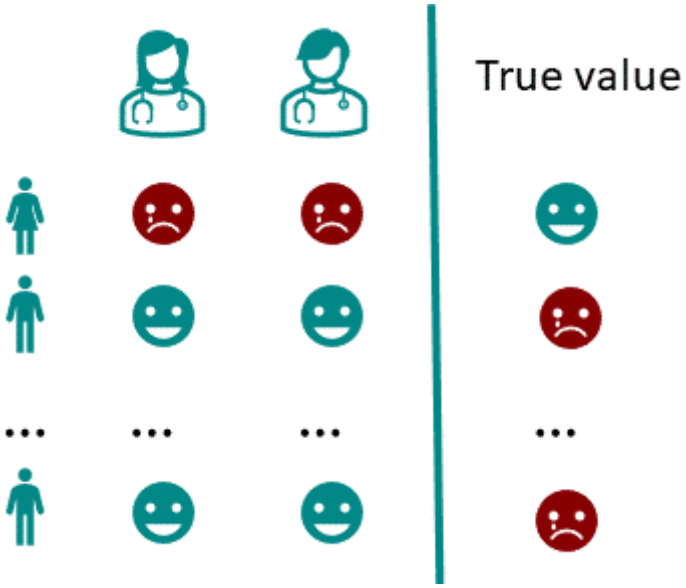


In this case, Cohen’s kappa score indicates how well the two measurements from the same person agree.

Measuring the agreement: Cohen's Kappa measures the agreement between two dependent categorical samples.

21.4 Cohen’s Kappa reliability and validity

It is important to note that the Cohen's Kappa coefficient can only tell you how reliably both raters are measuring the same thing. It does not tell you whether what the two raters are measuring is the right thing!



In the first case we speak of reliability (whether both are measuring the same thing) and in the second case we speak of validity (whether both are measuring the right thing). Cohen's Kappa can only be used to measure reliability.

21.5 Calculate Cohen’s Kappa

Now the question arises, how is Cohen's Kappa calculated? This is not difficult! We create a table with the frequencies of the corresponding answers.

For this we take our two raters, each of whom has rated whether a person is depressed or not. Now we count how often both have measured the same and how often not.

So, we make a table with Rater 1 with "not depressed" and "depressed" and Rater 2 with "not depressed" and "depressed". Now we simply keep a tally sheet and count how often each combination occurs.

Figure 129: Calculation of Cohen’s Kappa example

Let's say our final result is as follows: 17 people rated both raters as "not depressed." For 19 people, both chose the rating "depressed."

So, if both raters measured the same thing, that person is on the diagonal, if they measured something different, that person is on the edge. Now we want to know how often both raters agree and how often they don't.

Rater 1 and Rater 2 agree that 17 patients are not depressed and 19 are depressed. So both raters agree in 36 cases. In total, 50 people were assessed.

With these numbers, we can now calculate the probability that both raters are measuring the same thing in a person. We do this by dividing 36 by 50. This gives us the following result: In 72% of the cases, both raters assess the same, in 28% of the cases they rate it differently.

$$p_o = \frac{36}{50} = 72\%$$

This gives us the first part we need to calculate Cohen's Kappa. Cohen's Kappa is given by this formula:

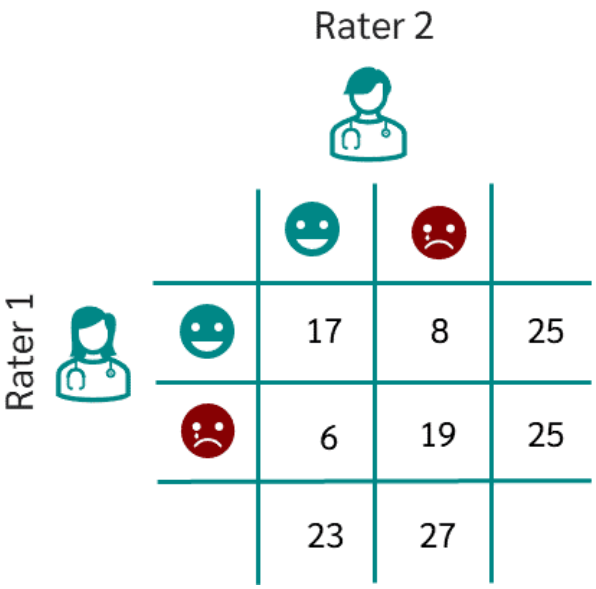
$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

So, we just calculated po, what is pe?

If both doctors were to answer the question of whether a person is depressed or not purely by chance, by simply tossing a coin, they would probably come to the same conclusion in some cases, purely by chance.

And that is exactly what p_e indicates: The hypothetical probability of a random match. But how do you calculate p_e?

To calculate p_e, we first need the sums of the rows and columns. Then we can calculate p_e.



$$p_e = \underbrace{\frac{25}{50} \cdot \frac{23}{50}}_{0.5 \cdot 0.46} + \underbrace{\frac{25}{50} \cdot \frac{27}{50}}_{0.5 \cdot 0.54}$$

$$0.23 + 0.27 = 0.5$$

In the first step, we calculate the probability that both raters would randomly arrive at the rating "not depressed."

- Rater 1 rated 25 out of 50 people as "not depressed", i.e. 50%.
- Rater 2 rated 23 out of 50 people as "not depressed", i.e. 46%.

The overall probability that both raters would say "not depressed" by chance is: 0.5 * 0.46 = 0.23

In the second step, we calculate the probability that the raters would both say "depressed" by chance.

- Rater 1 says "depressed" in 25 out of 50 persons, i.e. 50%.
- Rater 2 says "depressed" in 27 out of 50 people, i.e. 54%.

The total probability that both raters say "depressed" by chance is: 0.5 * 0.54 = 0.27. Now we can calculate p_e.

If both values are now added, we get the probability that the two raters coincidentally agree. p_e is therefore 0.23 + 0.27 which is equal to 0.50. Therefore, if the doctors had no guidance and simply rolled the dice, the probability of such a match is 50%.

Now we can calculate the Cohen's Kappa coefficient. We simply substitute p_o and p_e and we get a Kappa value of 0.4 in our example.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$= \frac{0.72 - 0.5}{1 - 0.5}$$

$$= 0.44$$

By the way, in p_o the o stands for "observed". And in p_e, the e stands for "expected". Therefore, p_o is what we actually observed and p_e is what we would expect if it were purely random.

21.6 Cohen's Kappa interpretation

Now, of course, we would like to interpret the calculated Cohens Kappa coefficient. The table of Landis & Koch (1977) can be used as a guide.

Kappa	
>0.8	Almost Perfect
>0.6	Substantial
>0.4	Moderate
>0.2	Fair
0-0,2	Slight
<0	Poor

Therefore, the calculated Cohen's Kappa coefficient of 0.44 indicates moderate reliability or agreement.

21.7 Cohen's Kappa Standard Error (SE)

The Standard Error (SE) of a statistic, like Cohen's Kappa, is a measure of the precision of the estimated value. It indicates the extent to which the calculated value would vary if the study were repeated multiple times on different samples from the same population. Therefore it is a measure of the variability or uncertainty around the Kappa statistic estimate.

21.8 Calculating Standard Error of Cohen's Kappa

The calculation of the SE for Cohen's Kappa involves somewhat complex formulas that account for the overall proportions of each category being rated and the distribution of ratings between the raters. The general formula for the SE of Cohen's Kappa is:

$$\mathrm{SE}(\kappa) = \sqrt{\frac{p_o(1-p_o)}{n(1-p_e)^2}}$$

Where *n* is the total number of items being rated.

21.9 Interpreting Standard Error

Small Standard Error: A small SE suggests that the sample estimate is likely to be close to the true population value. The smaller the SE, the more precise the estimate is considered to be.

Large Standard Error: A large SE indicates that there is more variability in the estimates from sample to sample and, therefore, less precision. It suggests that if the study were repeated, the resulting estimates could vary widely.

21.10 Calculate Cohen's Kappa with DATAtab

Now we will discuss how you can easily calculate Cohen's Kappa for your data online using DATAtab.

Simply go to datatab.net and copy your own data into the table. Now click on the tab "Reliability".

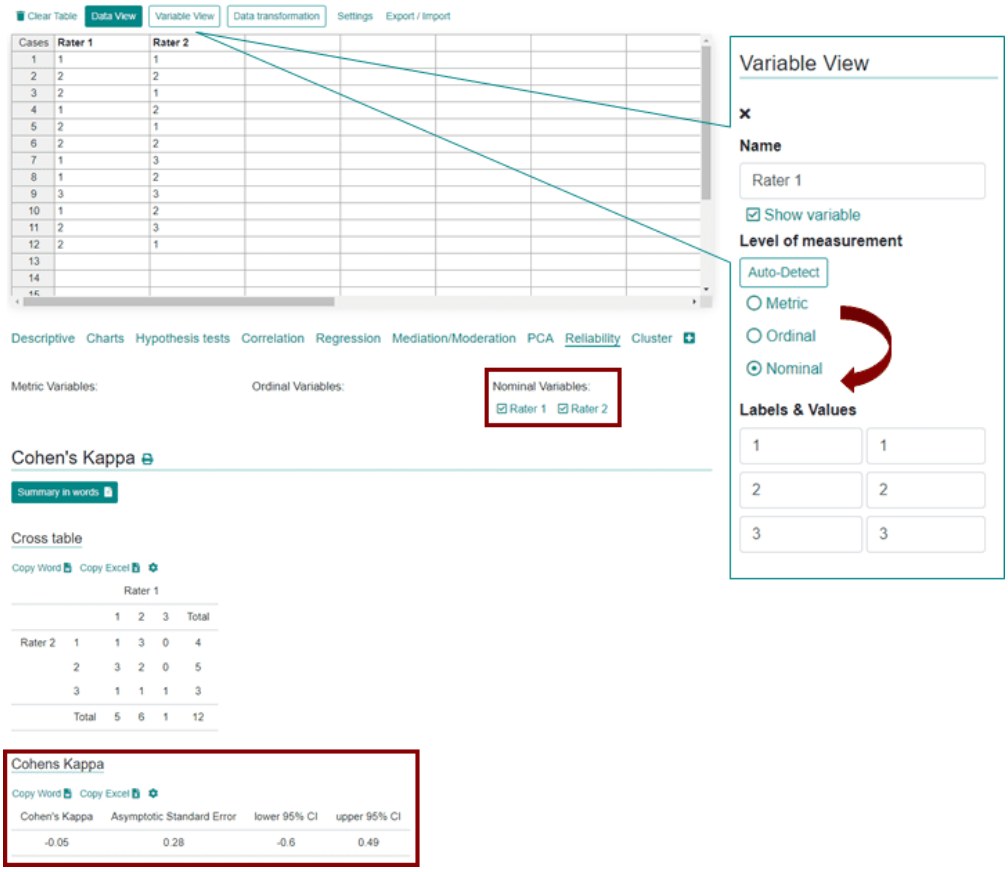


Figure 130: Cohen's Kappa calculation with DATAtab

All you have to do is click on the variables you want to analyse and Cohen's Kappa will be displayed automatically. First you will see the crosstab and then you can read the calculated Cohen's Kappa coefficient. If you don't know how to interpret the result, just click on interpretations in words.

An inter-rater reliability analysis was performed between the dependent samples Rater1 and Rater2. For this, Cohen's Kappa was calculated, which is a measure of the agreement between two related categorical samples. The Cohen's Kappa showed that there was moderate agreement between the samples Rater1 and Rater2 with $\kappa = 0.23$.

22. Weighted Cohen’s Kappa

Weighted Cohen's Kappa is a measure of the agreement between two ordinally scaled samples and is used whenever you want to know if two people's measurements agree. The two people who measure something are called raters.

In the case of a "normal" Cohen's Kappa, the variable to be measured by the two raters is a nominal variable. With a nominal variable, the characteristics can be distinguished, but there is no ranking between the characteristics.

Nominal

Characteristics can be **distinguished**

A D C B

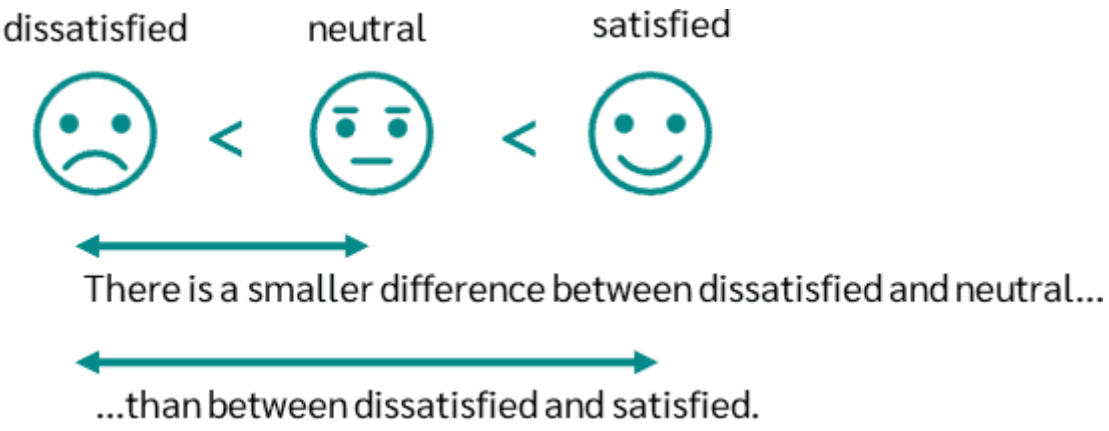
Ordinal

Characteristics can be **sorted**

A < B < C < D

Cohen's kappa takes into account whether the two raters measured the same thing or not, but it does not take into account the degree of disagreement. What if you don't have a nominal variable but an ordinal variable?

If you have an ordinal variable, that is, a variable in which the characteristics can be ordered, then of course you want to take that order into account.



Let's say your expressions are dissatisfied, neutral, and satisfied. There is a smaller difference between dissatisfied and neutral than between dissatisfied and satisfied. If you want to take the size of the difference into account, you have to use the weighted Cohen's Kappa.

So if you have a nominal variable, you use Cohen's Kappa. If you have an ordinal variable, you use the weighted Cohen's kappa.

22.1 Reliability and validity

It is important to note that the weighted Cohen's Kappa can only tell you how reliably both raters are measuring the same thing. It cannot tell you whether what the two raters are measuring is the right thing!

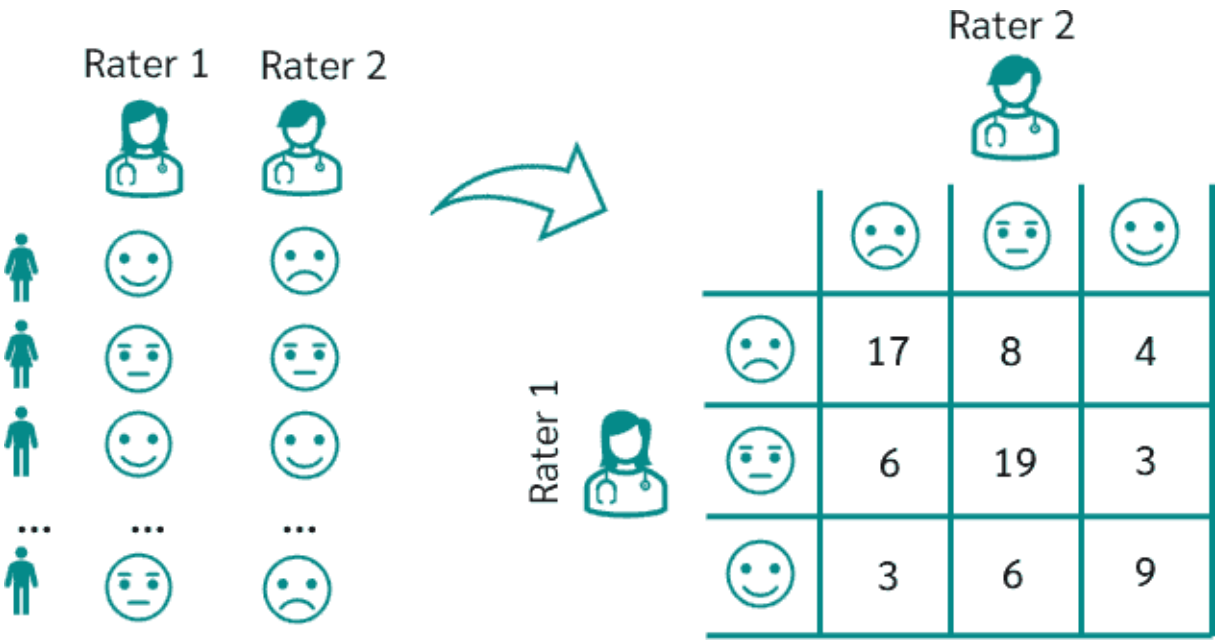
So if both raters are pretty much always measuring the same thing, you would have a very high weighted Cohen's Kappa. However, the weighted Cohen's Kappa does not tell you whether this measurement corresponds to reality, i.e. whether the raters are measuring the right thing! In the first case we are talking about reliability. In the second case we speak of validity.

22.2 Calculating weighted Cohen's Kappa

How is weighted Cohen's kappa calculated? Let's say two doctors have rated how satisfied they are with the therapeutic success of their patients. The doctors can answer with dissatisfied, neutral and satisfied.

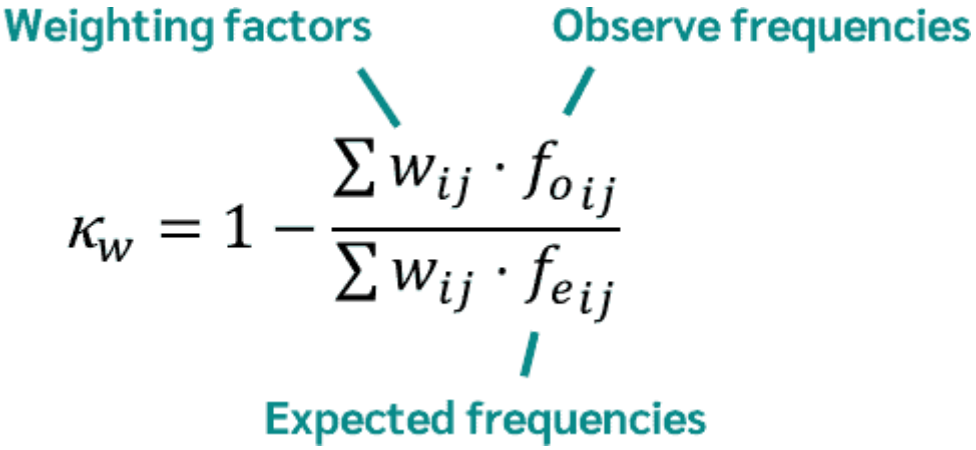
Now you want to know how much agreement there is between the two doctors. Since we have an ordinal variable with the rank order dissatisfied, neutral and satisfied, we determine the agreement with the weighted Cohen's kappa.

The first step is to create a table with the frequencies of each response. We plot one rater on each axis. Here we have our two raters, each of whom rated whether they were dissatisfied, neutral or satisfied with a person's success.



Let's say a total of 75 patients have been evaluated. Now let's count how often each combination occurs. Let's say 17 times both raters are dissatisfied, 8 times rater 1 is dissatisfied and rater 2 is neutral, 4 times rater 1 is dissatisfied and rater 2 is satisfied and so on and so forth. For the ratings on the diagonal, both raters agree.

Weighted Cohen's kappa can be calculated using the following formula:



Where w are the weighting factors, fo are the observed frequencies, and fe are the expected frequencies. Instead of the frequencies, we could also use the calculated probabilities, i.e. the observed probabilities po and the expected probabilities pe.

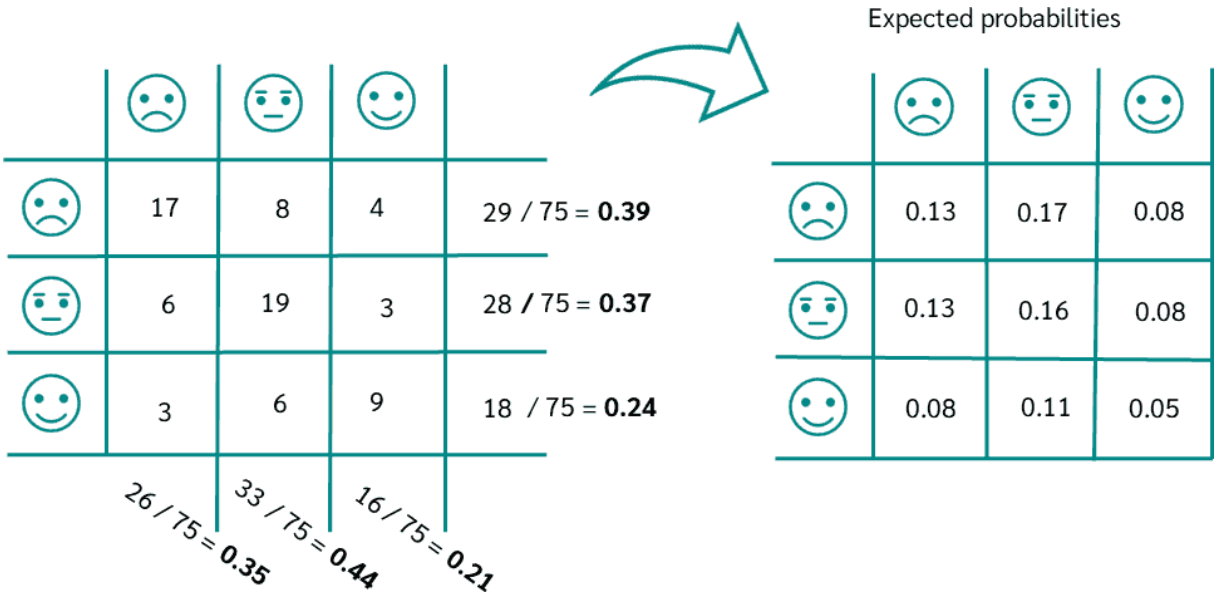
If we calculated Cohen's kappa using probabilities rather than frequencies, we would simply divide each frequency by the number of patients, i.e. 75, and have the observed probabilities.

But we still need the weights and the expected frequencies. Let's start with the expected frequencies.

22.3 Calculate expected frequency

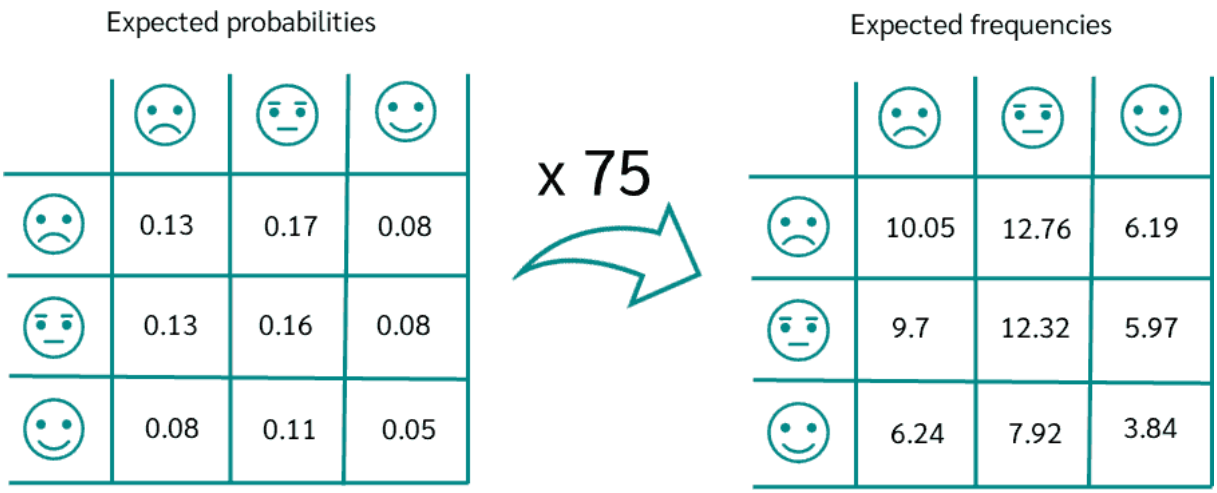
To calculate the expected frequency, we first calculate the sums of the rows and columns. So we simply add up all the rows and all the columns.

For example, in the first row we get a sum of 29 with 17 + 8 + 5. We now divide this by 75 of the total number of cases.



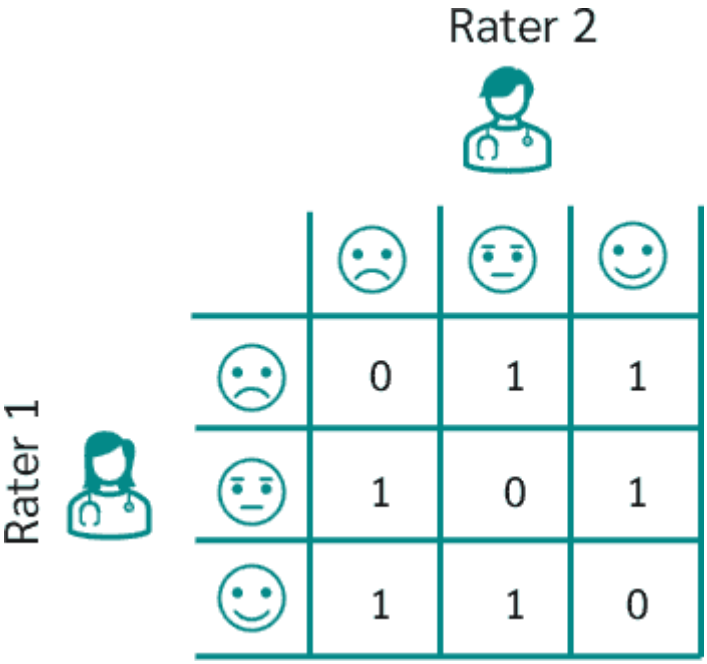
We can now calculate the expected probability for each cell by multiplying the row probability by the column probability. So for the first cell we get 0.35 times 0.39 which is 0.13, for the second cell we get 0.44 times 0.39 which is 0.17.

Now, if we multiply each probability by 75, we get the expected frequencies.



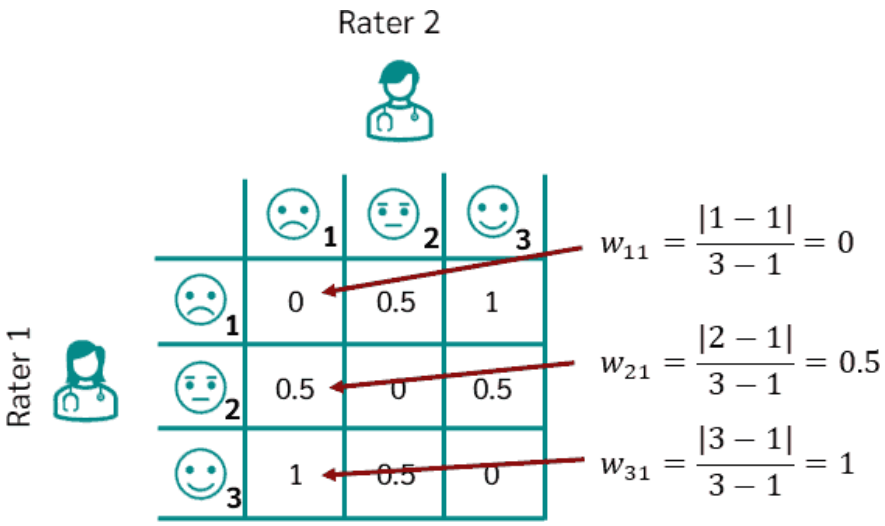
22.4 Calculate weighting matrix

If we did not use any weighting at all, our matrix would consist only of ones and zeros on the diagonal. If both raters gave the same answer, there would be a zero in the cell, otherwise there would be a one. It does not matter how far apart the raters are in their answers, if they answered something different it is weighted by 1.



The linear weighting matrix can be calculated using the following formula. Let i be the index for the rows and j for the columns. k is the number of expressions, in our case 3.

$$w_{ij} = \frac{|i-j|}{k-1}$$

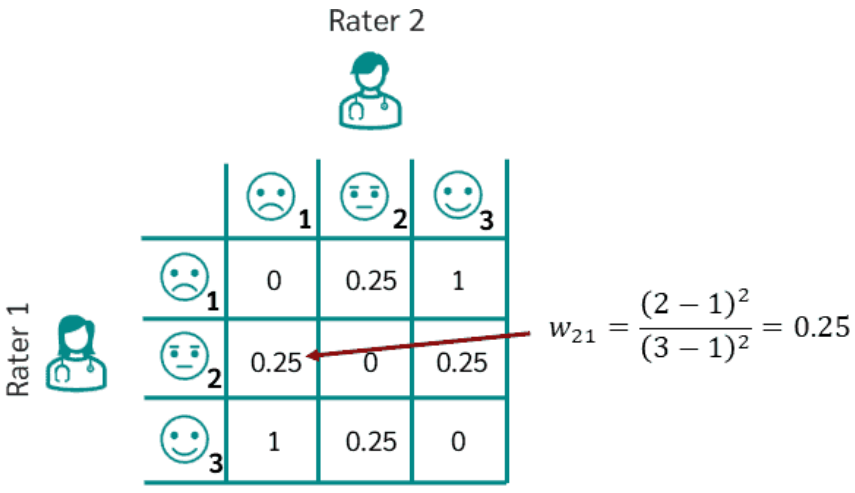


So now scores that are close together are weighted less than scores that are far apart.

22.5 Linear and quadratic weighting

What about quadratic weighting? If we use quadratic weighting instead of linear weighting, the distances are simply squared again. In this way, scores that are far apart are weighted even more heavily in relation to scores that are close together than in the linear case. The weighting matrix is then obtained with the following matrix.

$$w_{ij} = \frac{(i-j)^2}{(k-1)^2}$$



So we can now decide whether to use no weighting, linear weighting or quadratic weighting. We will continue with the linear weighting.

No weighting

	0	1	1
	1	0	1
	1	1	0

Linear weighting

	0	0.5	1
	0.5	0	0.5
	1	0.5	0

Quadratic weighting

	0	0.25	1
	0.25	0	0.25
	1	0.25	0

22.6 Calculate weighted Kappa

We can now calculate the weighted kappa. We have the weighting matrix, the observed frequency and the expected frequency. Let's start with the sum in the figure below. We simply multiply each cell of the weighting matrix by the corresponding cell of the observed frequency and add them up. So 0 times 17 + 0.5 times 8 to finally 0 times 9.

Weighting matrix

0	0,5	1
0,5	0	0,5
1	0,5	0

Observed frequency

17	8	4
6	19	3
3	6	9

$$\kappa_w = 1 - \frac{\sum w_{ij} \cdot p_{ij}}{\sum w_{ij} \cdot e_{ij}} = 1 - \frac{0 \cdot 17 + 0.5 \cdot 8 + \cdots + 0 \cdot 9}{0 \cdot 10.05 + 0.5 \cdot 12.76 + \cdots + 0 \cdot 3.84} = 1 - \frac{18.5}{30.6} = 0.396$$

0	0.5	1
0.5	0	0.5
1	0.5	0

Weighting matrix

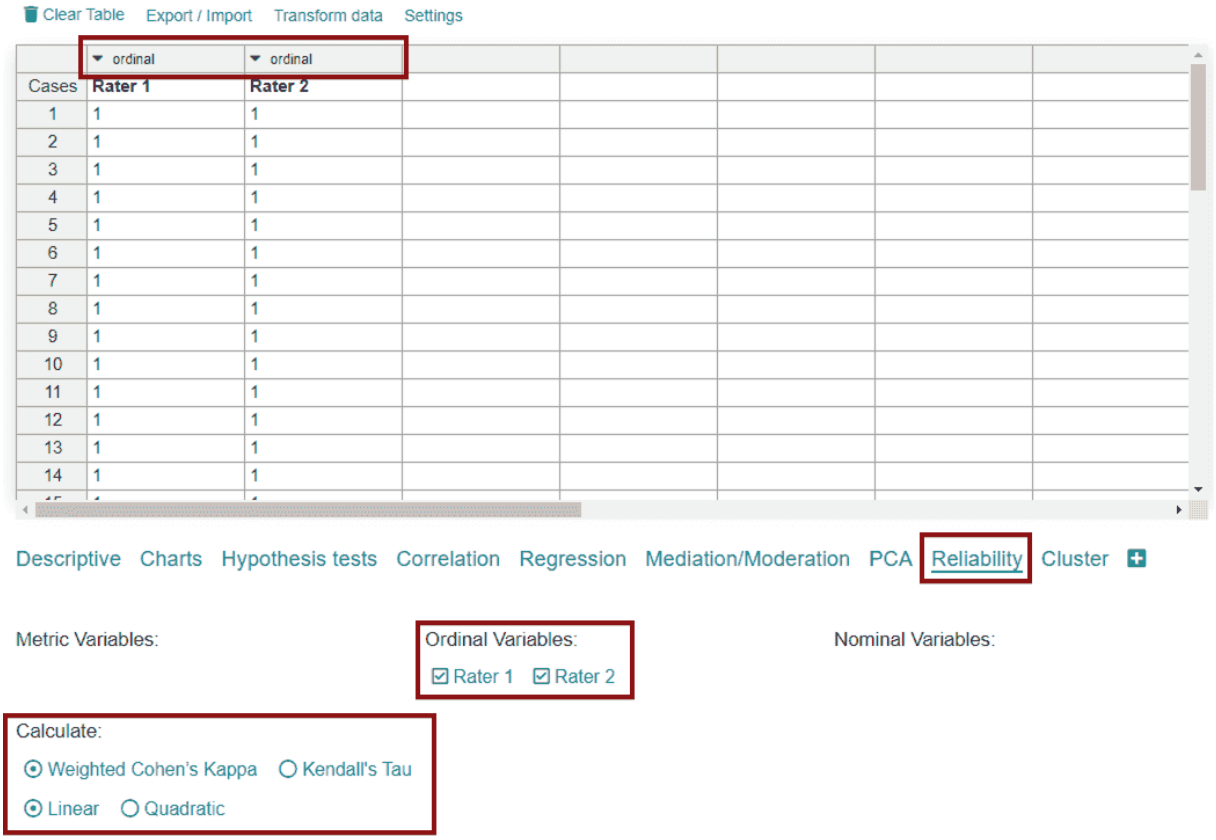
10.05	12.76	6.19
9.7	12.32	5.97
6.24	7.92	3.84

Expected frequency

We now do the same with the weighting matrix and the expected frequency. 0 times 10.05 plus 0.5 times 12.76 and finally 0 times 3.84. If we now calculate everything, we get a weighted kappa of 0.396.

22.7 Calculating weighted Cohen's Kappa with DATAtab

To calculate weighted Cohen's Kappa online, simply go to the Statistics Calculator, copy your own data into this table, and click on the Reliability tab.



DATAtab automatically tries to assign the appropriate scale level to the data, in this case DATAtab assumes that the data are nominal. If we clicked on Rater 1 and Rater 2, DATAtab would calculate the unweighted normal Cohen's kappa. However, in our case these are ordinal variables. So we simply change the scale level to ordinal.

Weighted Cohen's Kappa

Summary in words

Cross table

Copy Word Copy Excel

		Rater 1			
		1	2	3	Total
Rater 2	1	16	4	3	23
	2	6	10	0	16
	3	2	1	8	11
	Total	24	15	11	50

Weighted Cohen's Kappa

Copy Word Copy Excel

Weighted Cohen's Kappa	Standard Error	lower 95% CI	upper 95% CI	p
0.5	0.11	0.28	0.72	<.001

If we now click on both raters, the weighted Cohen's kappa is calculated. We can now choose whether we want linear or quadratic weighting. Here we see the cross table, which shows us how often each combination occurs. Then we get the results for the Cohen's kappa. With this data we get a weighted Cohen's kappa of 0.05.

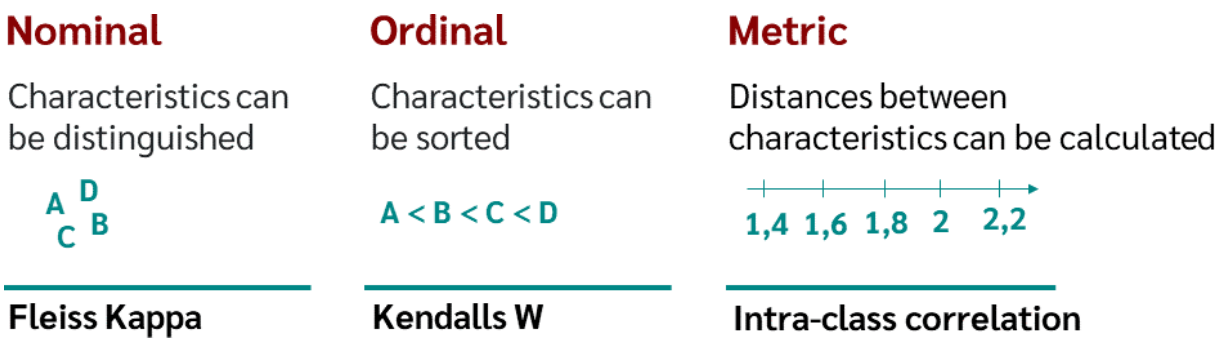
If you're not sure how to interpret the results, you can click on Summary in Words: An inter-rater reliability analysis was performed between the dependent samples Rater1 and Rater2. This was done by calculating the Weighted Cohens Kappa, which is a measure of the agreement between two related categorical samples. The Weighted Cohens Kappa showed that there was moderate agreement between the Rater1 and Rater2 samples with κ= 0.5.

23. Fleiss Kappa

You use Fleiss Kappa whenever you want to know if the measurements of more than two people agree. The people who measure something are called raters.

In the case of the Fleiss Kappa, the variable to be measured by the three or more rates is a nominal variable. Therefore, if you have a nominal variable, you use the Fleiss Kappa.

If you had an ordinal variable and more than two raters, you would use the Kendall's W and if you had a metric variable, you would use the intra-class correlation. If you had only two raters and a nominal variable, you would use Cohen's Kappa.



But that's enough theory for now, let's look at an example.

23.1 Fleiss Kappa Example

Let's say you have developed a measuring instrument, for example a questionnaire, that doctors can use to determine whether a person is depressed or not.

Now you give the measuring instrument to doctors and let them assess 50 people with it. The big question is: how well do the doctors' measurements agree?

If the ratings of the raters agree very well, the inter-rater reliability is high.

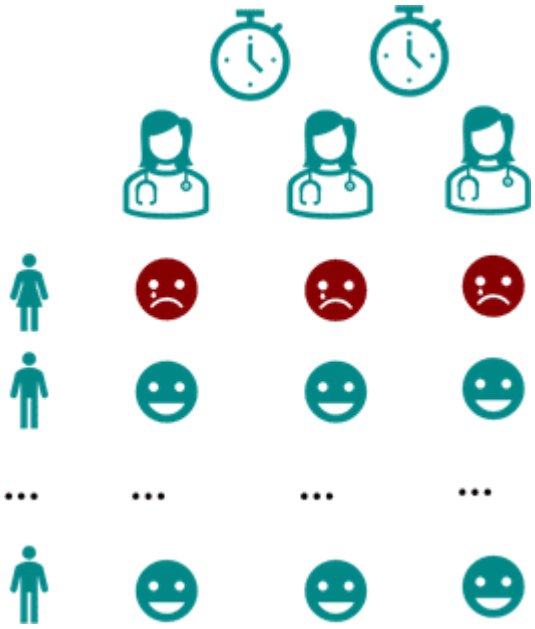
And it is this inter-rater reliability that is measured by Fleiss Kappa. Fleiss Kappa is a measure of inter-rater reliability.

Definition: The Fleiss Kappa is a measure of how reliably three or more raters measure the same thing.

23.2 Fleiss Kappa with repeated measurement

So far, we have considered the case where two or more people measure the same thing. However, Fleiss Kappa can also be used when the same rater makes the measurement at more than two different times.

In this case, Fleiss Kappa indicates how well the measurements of the same person match.

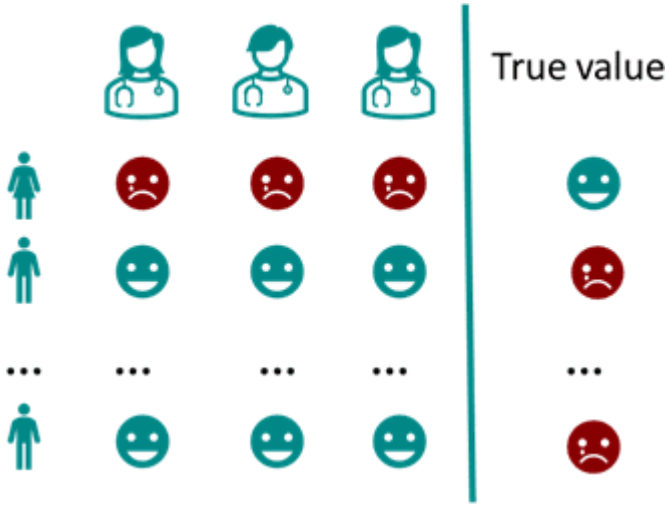


In this case, the variable of interest has two expressions, depressed and non-depressed; of course, the variable of interest may consist of more than two expressions.

Measure of the agreement: Fleiss Kappa is a measure of the agreement between more than two dependent categorical samples.

23.3 Fleiss Kappa reliability and validity

It is important to note that Fleiss Kappa can only tell you how reliably the raters are measuring the same thing. It cannot tell you whether what the raters are measuring is the right thing!

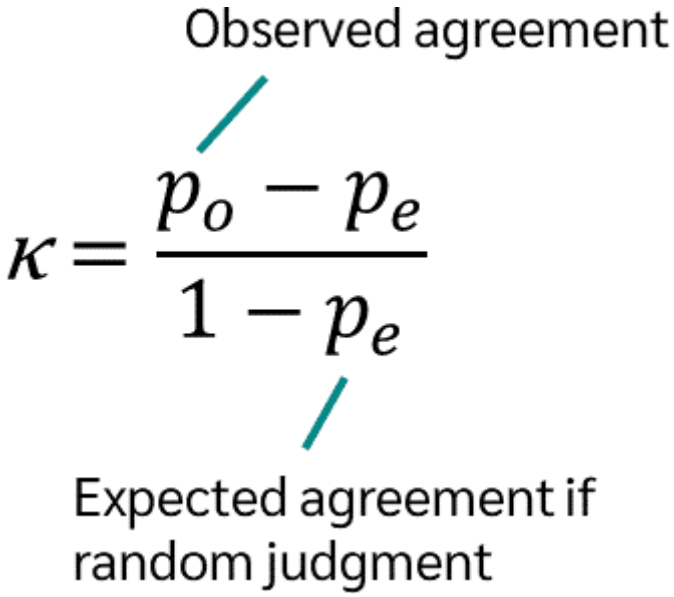


So if all the raters measured the same thing, you would have a very high Fleiss Kappa. Fleiss Kappa does not tell you whether this measured value corresponds to reality, i.e. whether the correct value is measured!

In the first case we speak of reliability, in the second of validity.

23.4 Calculate Fleiss Kappa

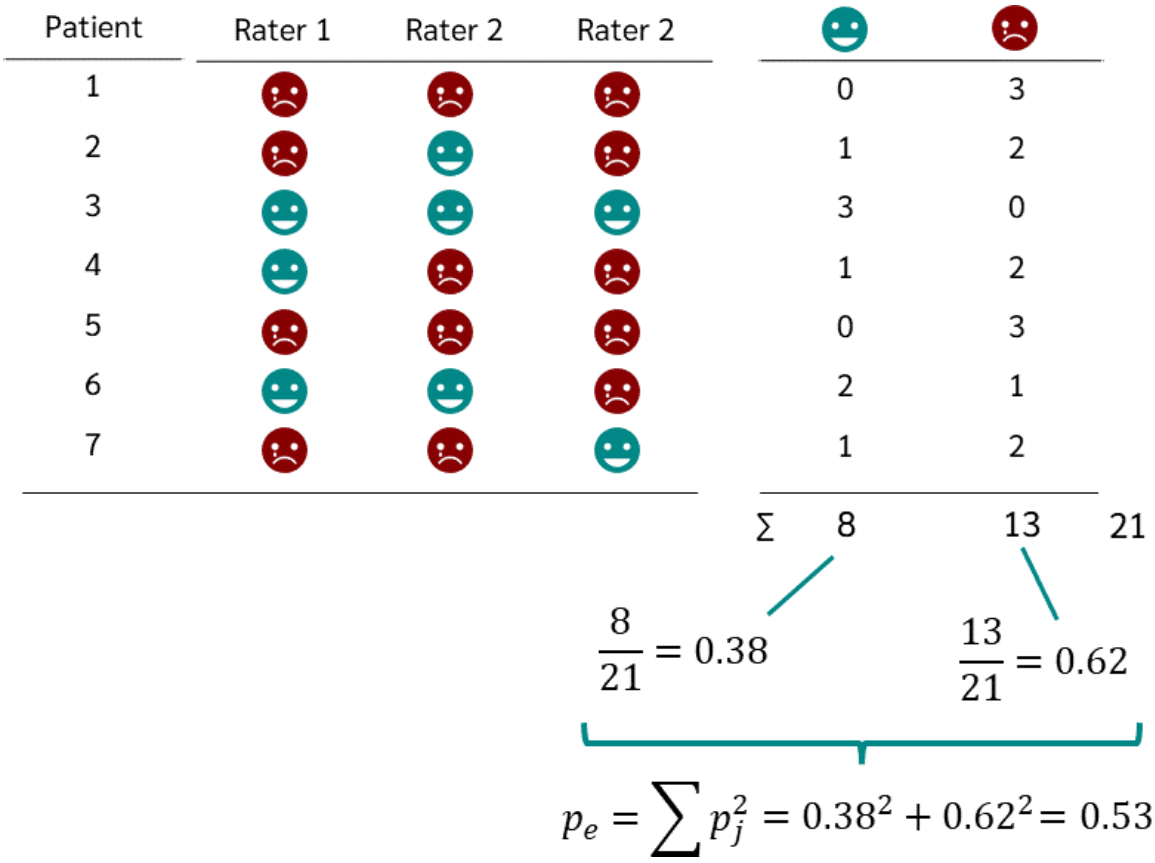
With this equation we can calculate the Fleiss Kappa:



In this equation, p_o is the observed agreement of the raters and p_e is the expected agreement of the raters. The expected agreement is given if the raters judge completely randomly, i.e. simply flip a coin for each patient to see whether they are depressed or not.

So how do we calculate p_o and p_e? Let's start with p_e Let's say we have 7 patients and three raters. Each patient has been assessed by each rater.

In the first step, we simply count how many times a patient was judged to be depressed and how many times they were judged not to be depressed.



For the first patient, 0 raters said that this person is not depressed and 3 raters said that this person is depressed. For the second person, one rater said that the person is not depressed and two said that the person is depressed.

Now we do the same for all the other patients and we can calculate the total for each one. In total we have 8 ratings with not depressed and 13 ratings with depressed. In total there were 21 ratings.

This allows us to calculate how likely a person is to be rated as not depressed or as depressed. To do this, we divide the number of ratings of depressed and not depressed by the total number of 21.

So we divide 8 by 21 to get 38% of the patients rated as not depressed by the raters and then we divide 13 by 21 to get 62% of the patients rated as depressed.

To calculate pe, we now square and sum the two values. So 0.382 plus 0.622 is 0.53.

Now we need to calculate p_o. p_o we can calculate with the following formula, don't worry, it looks more complicated than it is.

Patient	Rater 1	Rater 2	Rater 2	😀	☹
1	☹	☹	☹	0	3
2	☹	😀	☹	1	2
3	😀	😀	😀	3	0
4	😀	☹	☹	1	2
5	☹	☹	☹	0	3
6	😀	😀	☹	2	1
7	☹	☹	😀	1	2

$$p_0 = \frac{1}{N \cdot n \cdot (n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - N \cdot n \right)$$

$$\frac{1}{7 \cdot 3 \cdot (3-1)} = 0.024 \qquad 0^2 + 3^2 + \cdots 1^2 + 2^2 = 47 \qquad 7 \cdot 3 = 21$$

$$p_0 = 0.024 \cdot (47 - 21) = 0.624$$

Let's start with the first part. Capital *N* is the number of patients, so 7, and small n is the number of raters, so 3. This gives us 0.024 for the first part.

In the second part of the formula, we simply square each value in the table and add them up. So 0^2 plus 32 to finally 12 plus 22. This gives us 47.

And the third part is 7 times 3, which is 21. If we insert everything, we get 0.024 times 47 - 21, which is equal to 0.624.

So now we have p_o and p_e. Putting them into the equation for Kappa, we get a kappa of 0.19.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$= \frac{0.624 - 0.53}{1 - 0.53} = 0.19$$

23.5 Fleiss Kappa interpretation

Now, of course, the Fleiss Kappa coefficient must be interpreted. For this we can use the table from Landis and Kock (1977).

Kappa	Level of Agreement
> 0,8	Almost perfect
> 0,6	Substantial
> 0,4	Moderate
> 0,2	Fair
> 0	Slight
< 0	No agreement

Landis & Koch (1977)

For a Fleiss Kappa value of 0.19, we get just a slight match.

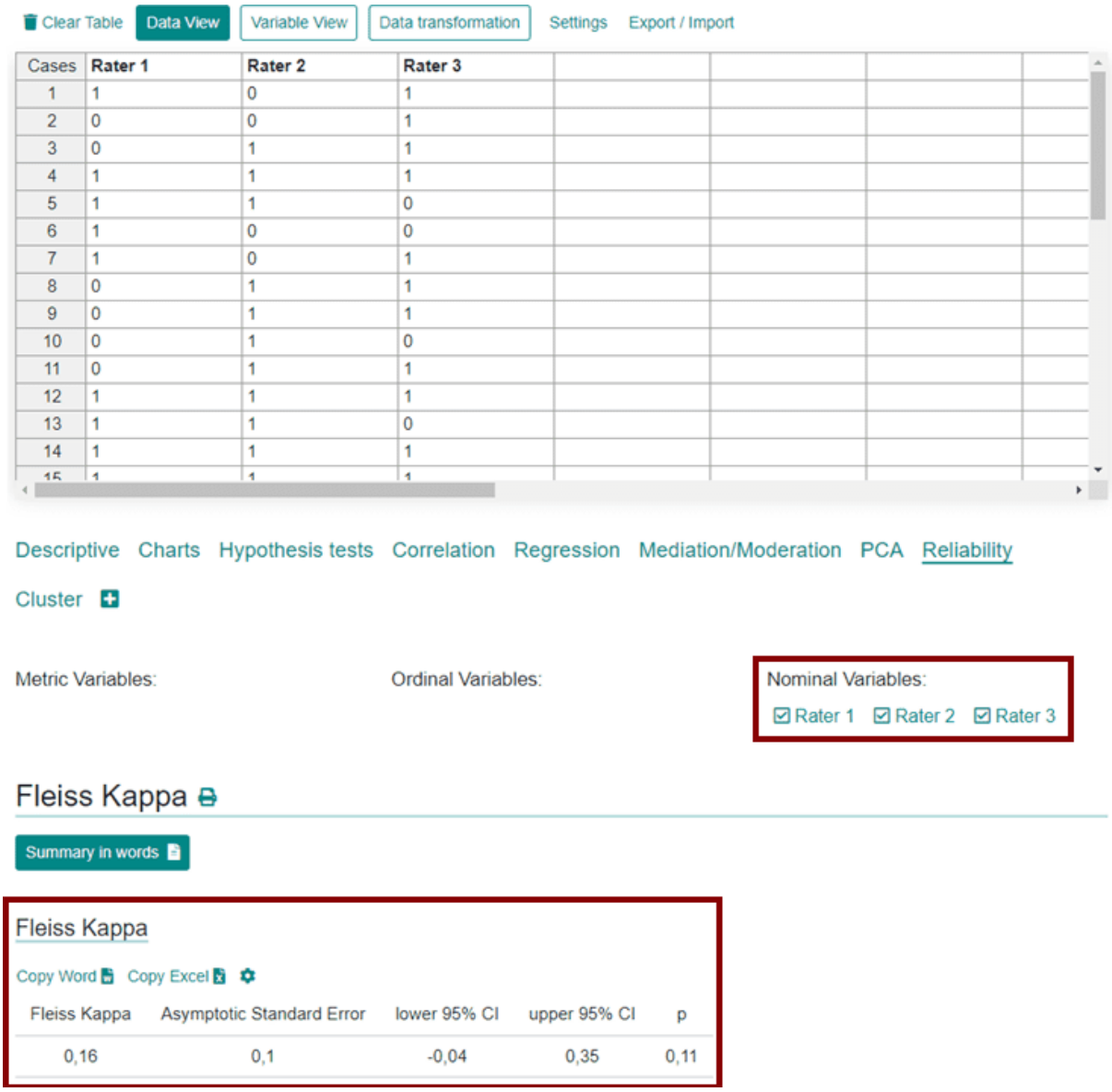
23.6 Calculate Fleiss Kappa with DATAtab

With DATAtab you can easily calculate the Fleiss Kappa online. Simply go to datatab.net and copy your own data into the table at the Fleiss Kappa calculator. Now click on the Reliability tab. Under Reliability you can calculate different reliability statistics, depending on how many variables you click on and which scale level they have, you will get a suitable suggestion.

The Fleiss Kappa is calculated for nominal variables. If your data is recognised as metric, please change the scale level under Data View to nominal.

If you now click on Rater 1 and Rater 2, the Cohen's Kappa will be calculated, if you now click on Rater 3, the Fleiss Kappa will be calculated.

Below you can see the calculated Fleiss Kappa.



Cases	Rater 1	Rater 2	Rater 3
1	1	0	1
2	0	0	1
3	0	1	1
4	1	1	1
5	1	1	0
6	1	0	0
7	1	0	1
8	0	1	1
9	0	1	1
10	0	1	0
11	0	1	1
12	1	1	1
13	1	1	0
14	1	1	1

Figure 131: Calculating Fleiss Kappa with DATAtab

If you don't know how to interpret the result, just click on Interpretations in Words.

An inter-rater reliability analysis was performed between the dependent samples of Rater 1, Rater 2 and Rater 3. For this purpose, the Fleiss Kappa was calculated, which is a measure of the agreement between more than two dependent categorical samples.

The Fleiss Kappa showed that there was a slight agreement between the samples of Rater 1, Rater 2 and Rater 3 with $\kappa = 0.16$.

24. Survival time analysis

This chapter is about Survival time analysis. We start with the question what a survival analysis is, then we come to the important point what the censoring of data means and then we briefly discuss the Kaplan Maier curve, the log rank test, and the Cox regression.

24.1 Basics of survival time analysis

Survival time analysis is a group of statistical methods in which the variable under study is the time until an event occurs. What does "time to occurrence of an event" mean?

Figure 132: Survival time analysis

Survival time analysis considers a variable that has a start time and, when a particular event occurs, an end time. The time between the start time and the event is the focus of survival analysis. For example, time may be measured in days, weeks or months.

24.2 Use cases for survival time analysis

An example would be to consider the time between a drug withdrawal and the relapse of the respective person. The start time would then be the end of withdrawal and the event under consideration would then be relapse. For example, you might be interested in whether different types of treatment have an impact on the time to relapse.

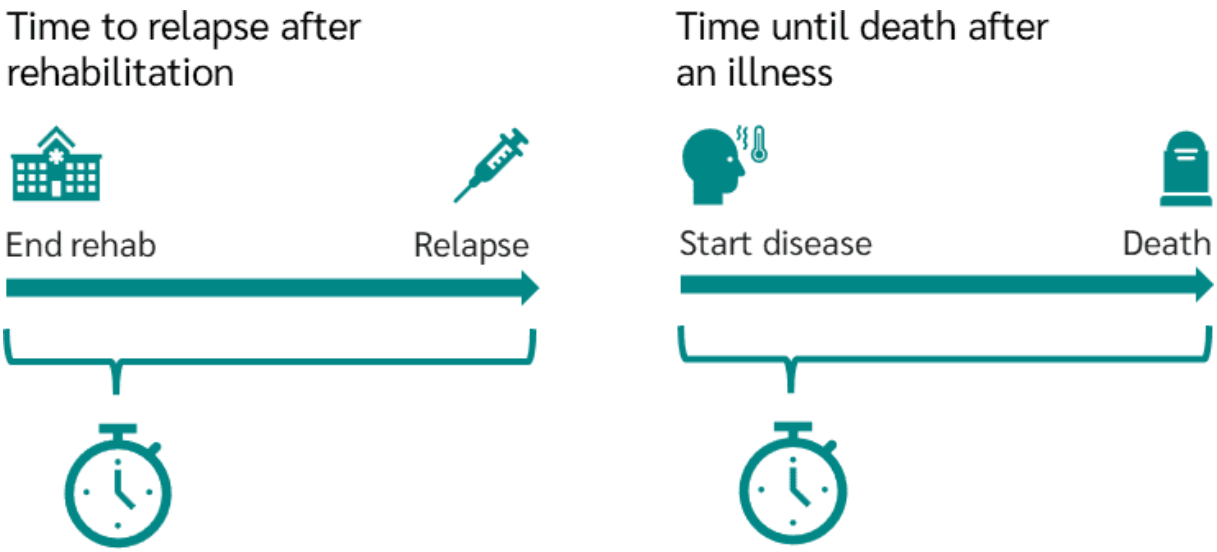
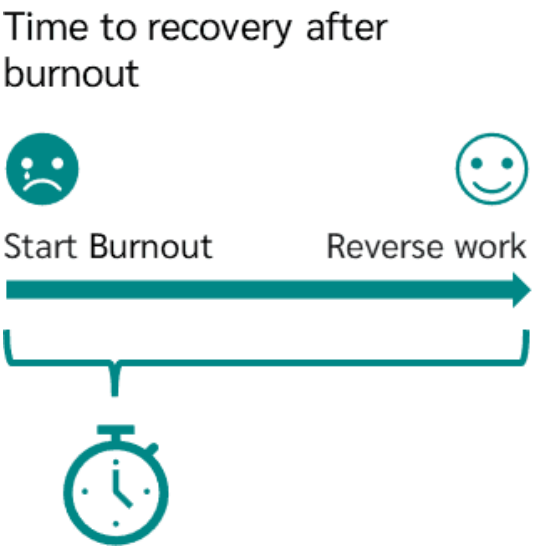


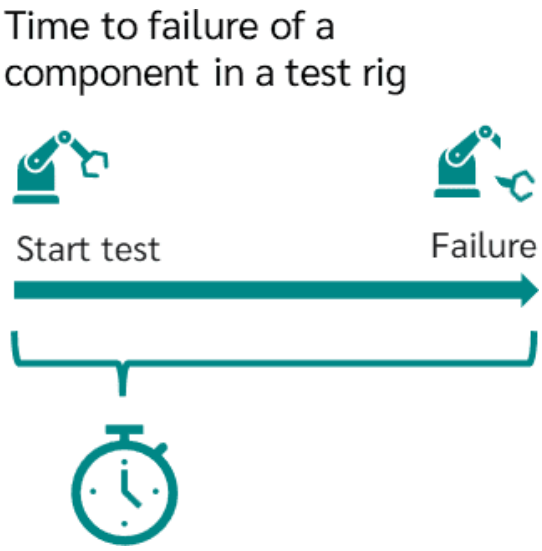
Figure 133: Survival time analysis use case

As the name "survival time analysis" implies, there is also a classic example: the time until death after a disease. Here, the start time is the recognition of the disease and the end time is death. Of great interest then is often whether a certain drug has an influence on the survival time.

Of course, the event doesn't have to be a dramatic event such as death, you could also look at the time to return to work after a burnout for example.



Moreover, the object under investigation need not be a human being. In engineering, for example, a common question is how long a component will last in a test without failing. In this case, different parameters could be varied to see if they have an effect on the object's survival time.



The time considered must have nothing to do with the actual "survival time", nevertheless one speaks of the survival time and the survival time analysis. The next question is how exactly a survival time analysis is performed. We will now take a look at an example.

24.3 Example of survival time analysis

How exactly is a survival analysis performed? Let us look at an example. Let's say you are a dental technician and you want to analyse the "survival time" of a filling in a tooth.

So your start time is the moment a person goes to the dentist for a filling. The end time, or event, is the moment when the filling breaks out. You are now interested in the time between these two events.

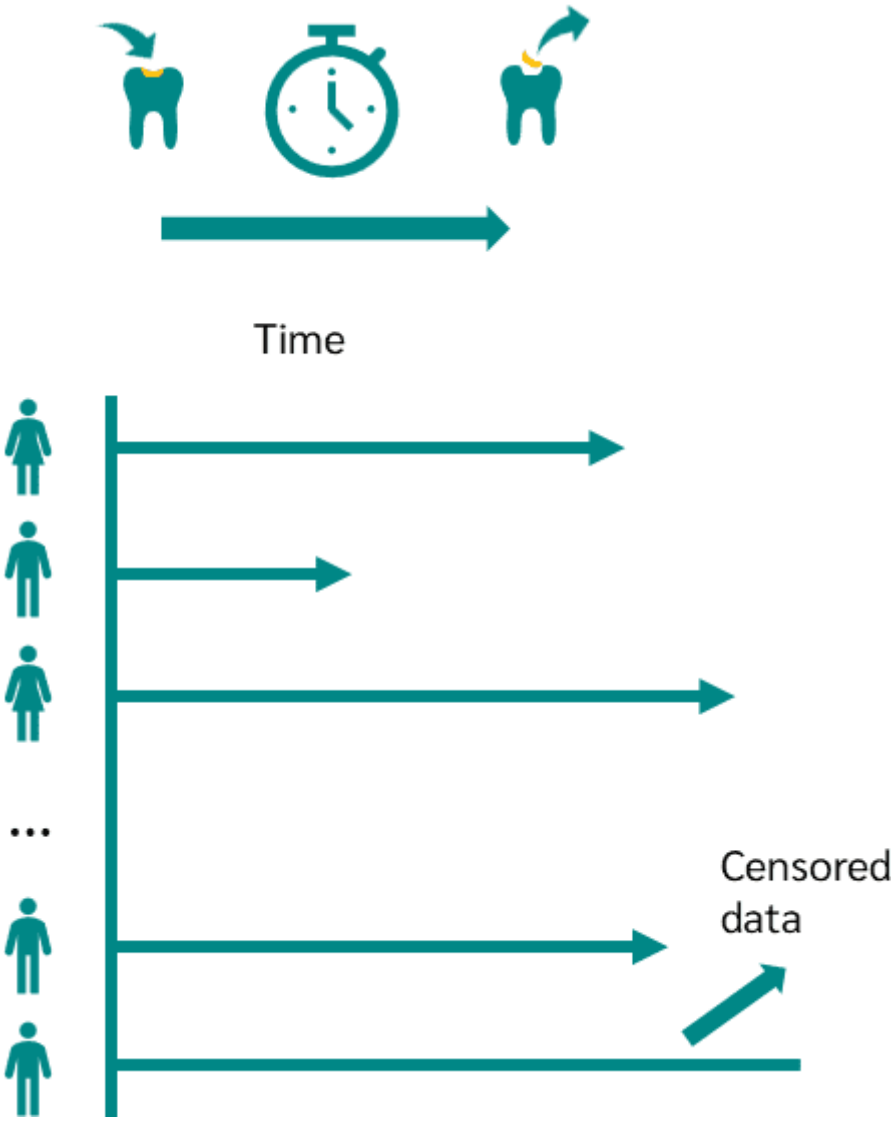


Figure 134: Survival time analysis example

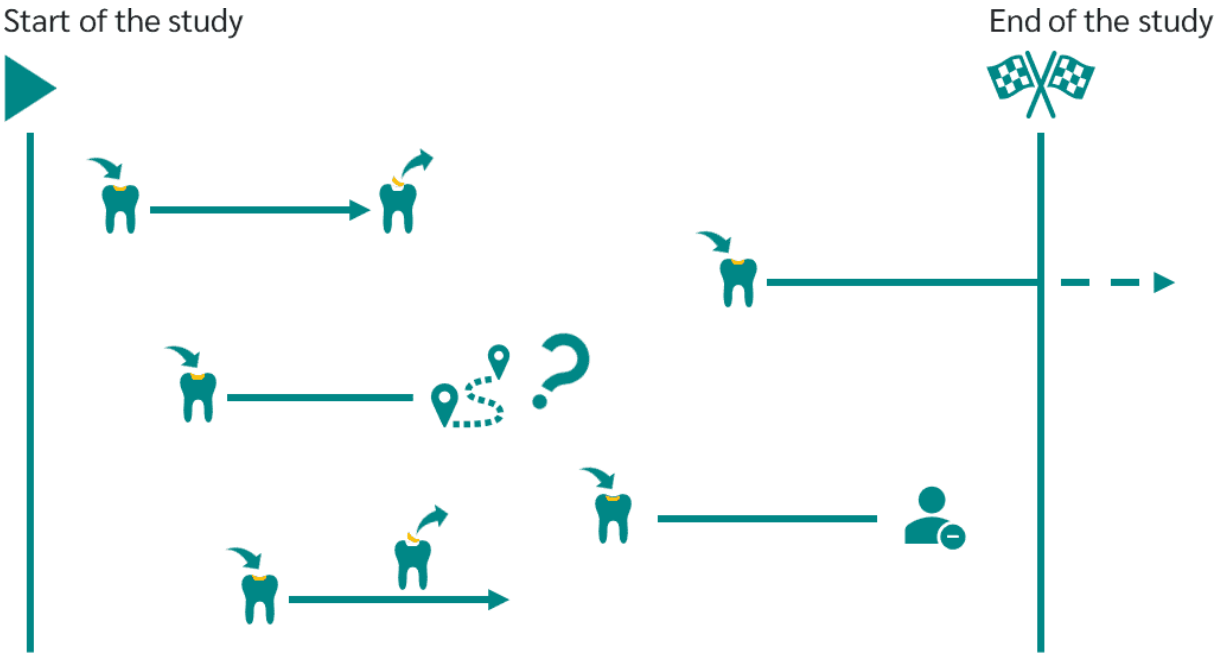
First, of course, you need test subjects so that you have data that you can evaluate. For each subject, you can now note the time that elapses until the filling breaks out.

Now you're probably asking yourself the question: What if a test person's tooth filling doesn't break out at all? Or what happens if a person moves, changes dentists and it is simply not known when the filling will break out?

All these cases are summarized under the term "censoring". Now let's look at what exactly is meant by this.

24.4 Censored data

First of all, it is important to keep in mind that a study cannot last indefinitely but extends over a limited period of time. For resource reasons (time, financial, etc.) and simply because you want to publish the results at some point, every study has a clear start and end date.



If a filling is inserted within this period and then the filling also breaks out again within this period and this is also documented, a valid case exists. The event has occurred.

However, it is also possible that a filling is inserted and then the end of the study is reached before the event occurs. Or it can happen that a subject decides not to continue with the study. In both cases, you do not know when or if the event under consideration has occurred.

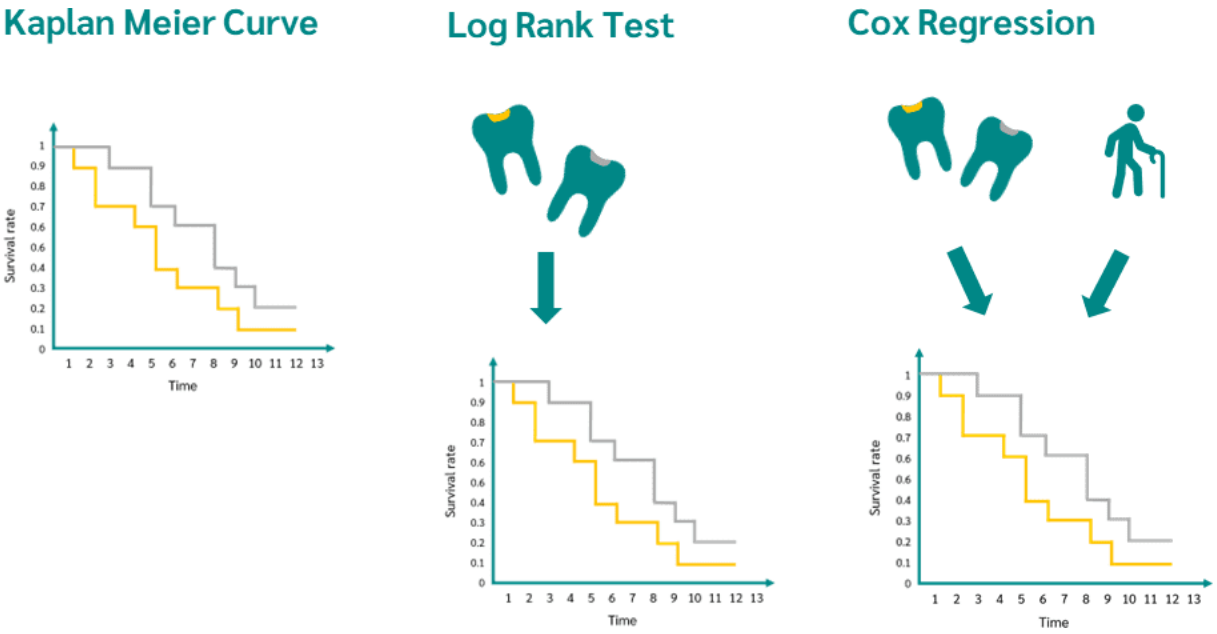
Further another event can occurs, that is not considered in the study. For example, the patient could die or even lose the whole tooth. In both cases, the event considered, that the filling breaks out, can no longer occur.

It can also happen that the patient does not notice that the filling has broken out and it is only discovered at the next routine check-up.

All in all, there are many cases where data is not fully available. This data is called "censored data". You will learn how to deal with this data in the Kaplan-Meier curve tutorial. Now let's look at the most common methods of survival analysis.

24.5 Methods of survival time analysis

The three most common methods of survival time analysis are (1) Kaplan Meier survival time curves, (2) the log rank test, and (3) Cox regression.



We will now briefly cover all three of these areas, and then I will show you how to easily calculate these methods online using DATAtab. For each of the three methods there is a detailed separate tutorial with calculation examples.

24.6 Kaplan-Meier Curve

The Kaplan-Meier curve is commonly used to analyze time-to-event data, such as the time until death or the time until a specific event occurs.

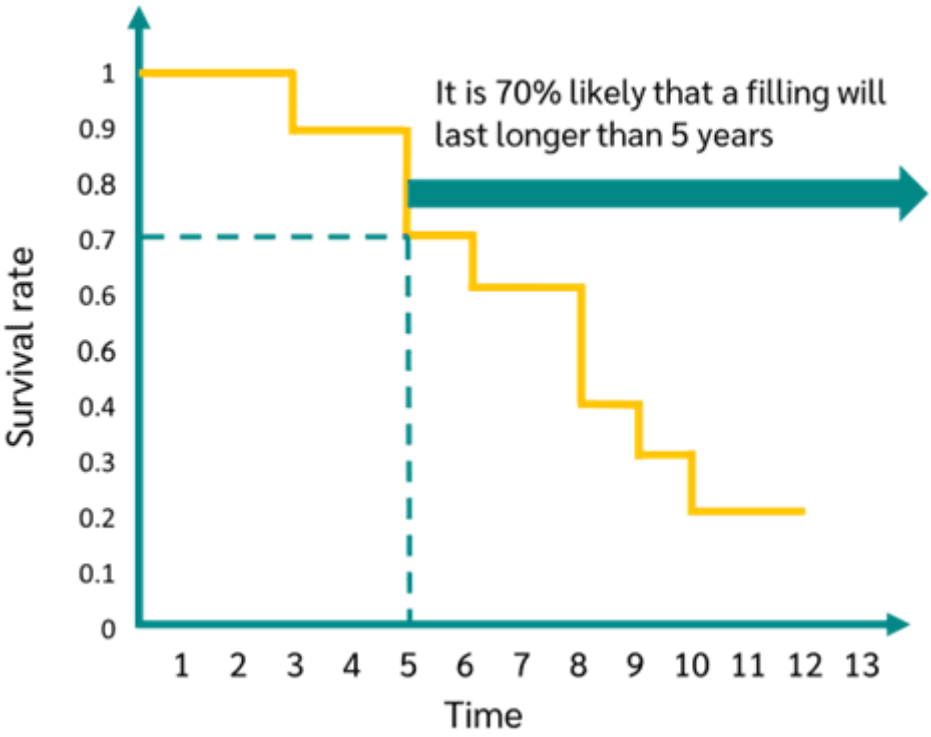
For this, the Kaplan Meier curve graphically represent the survival rate or survival function. Time is plotted on the x-axis and the survival rate is plotted on the y-axis.

24.6.1 Survival rate

The first question is what is the survival rate. Let's look at this with an example. Suppose you're a dental technician and you want to study the "survival time" of a filling in a tooth.

So your start time is the moment when a person goes to the dentist for a filling, and your end time, the event, is the moment when the filling breaks. The time between these two events is the focus of your study.

You can now see how likely it is that a filling will last longer than a certain point in time by looking at the Kaplan-Meier curve.



Thus the horizontal axis represents time, usually measured in months or years. The vertical axis represents the estimated probability.

For example, you may be interested in the probability that your filling will last longer than 5 years. To do this, you read off the value at 5 years on the graph, which is the survival rate. At 5 years, the Kaplan-Meier curve gives you a value of 0.7. So there is a 70% chance that your filling will last longer than 5 years.

24.6.2 Interpreting the Kaplan-Meier curve

The Kaplan-Meier curve shows the cumulative survival probabilities.

A steeper slope indicates a higher event rate (death rate) and therefore a worse survival prognosis. A flatter slope indicates a lower event rate and therefore a better survival prognosis. The curve may have plateaus or flat areas, indicating periods of relatively stable survival.

If there are multiple curves representing different groups, you can compare their shapes and patterns. If the curves are parallel, it suggests that the groups have similar survival experiences. If the curves diverge or cross, it indicates differences in survival between the groups.

At specific time points, you can estimate the survival probability by locating the time point on the horizontal axis and dropping a vertical line to the curve. Then, read the corresponding survival probability from the vertical axis.

24.6.3 Calculating the Kaplan-Meier curve

To create a Kaplan-Meier curve, you first need the data for your subjects. Let's say the filling lasted 3 years for the first subject, 4 years for the second subject, 4 years for the third subject, and so on.

Subject	Time
1	3
2	4
3	4
4	6
5	7
6	7
7	8
8	10
9	11
10	13

...

Time	m	n	S(t)
0	0	10	10/10 = 1
3	1	10	9/10 = 0.9
4	2	9	7/10 = 0.7
6	1	7	6/10 = 0.6
7	2	6	4/10 = 0.4
8	1	4	3/10 = 0.3
10	1	3	2/10 = 0.2
11	1	2	1/10 = 0.1
13	1	1	0/10 = 0

Let's assume that none of the cases are "censored". The data are already arranged so that the shortest survival time is at the top and the longest at the bottom.

Now we create a second table that we can use to draw the Kaplan-Meier curve. To do this, we look at the time points in the left table and add the time zero. So we have the time points 0, then 3, 4, 6, 7, 8 11 and 13. In total we have 10 subjects.

Now we look at how many fills break out at each time. We enter this in the column m. So at time 0, no fillings were broken out. After 3 years, there were no broken fillings, after 4 years there were two, after 6 years there was one. We now do the same for all the other times.

Next, we look at the number of cases that have survived to the time plus the number of cases where the event occurs at the exact time. We enter this in column n.

So n is the number of cases that survived to that point, plus the people who dropped out at that exact point.

After zero years we still have all 10 people. After 3 years, we get 10 for n, 9 people still have their fill intact, and one person's fill broke out exactly after 3 years.

The easiest way to get n is to take the previous n value and subtract the previous m value. So we get 10 - 1 equals 9. Then 9 minus 2 equals 7, 7 - 1 equals 6... and so on and so forth.

From column n we can now calculate the survival rates. To do this, we simply divide n by the total number, i.e. 10.

So 10 divided by 10 is equal to 1, 9 divided by 10 is equal to 0.9, 7 divided by 10 is equal to 0.7. Now we do the same for all the others.

24.6.4 Drawing Kaplan Meier curve

We can now plot the Kaplan-Meier curve. At time 0 we have a value of 1, after 3 years we have a value of 0.9 or 90%. After 4 years we get 0.7, after 6 years 0.6 and so on and so forth.

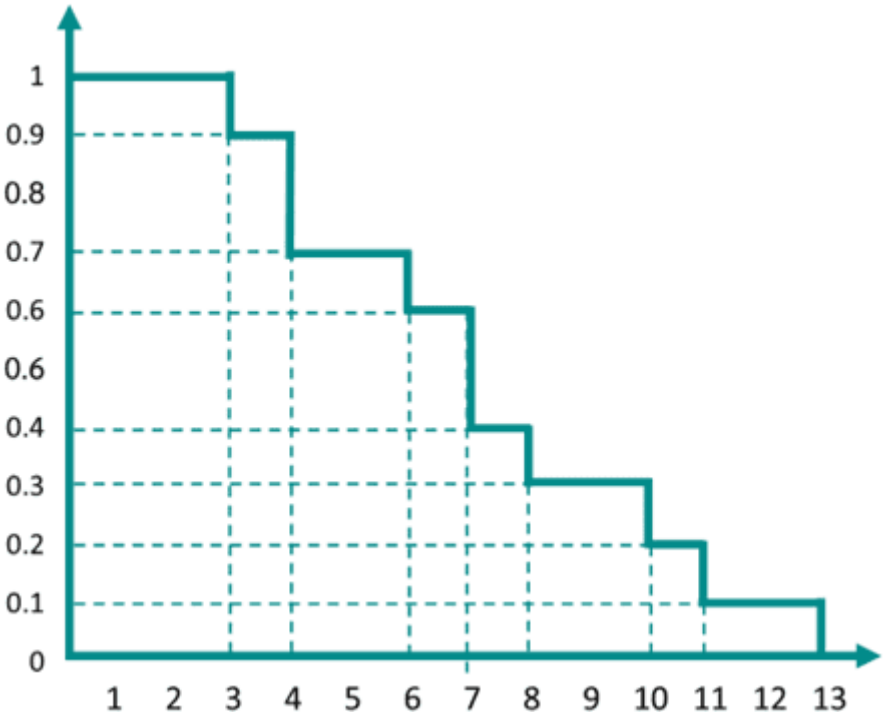
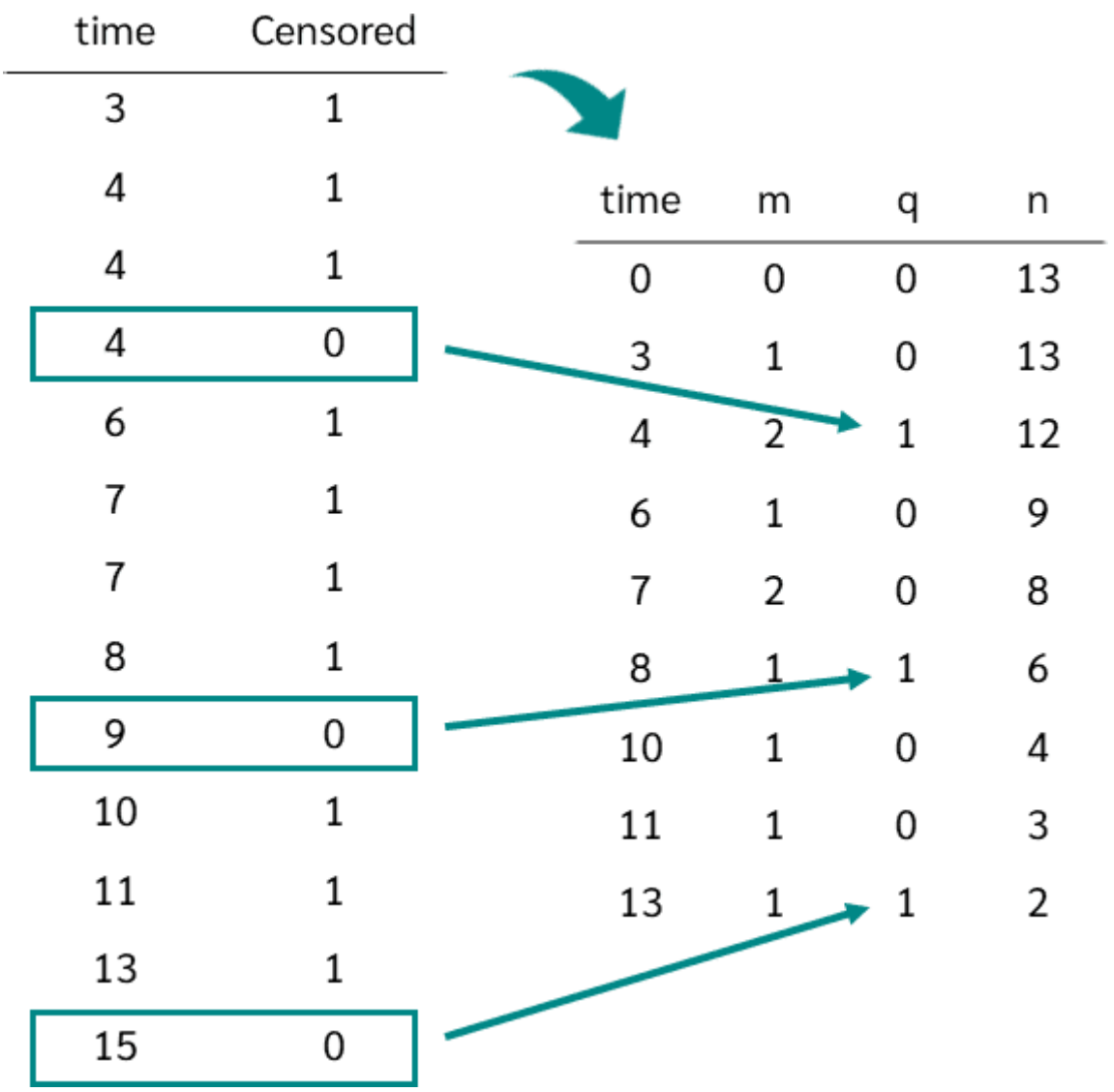


Figure 135: Kaplan Meier Curve

From the Kaplan-Meier curve, we can now see what percentage of the filling has not broken out after a certain time.

24.6.5 Censored data

Next, we look at what to do when censored data is present. For this purpose, censored data has been added to the example in these three places. If you're not sure what censored data is, see the survival analysis tutorial.



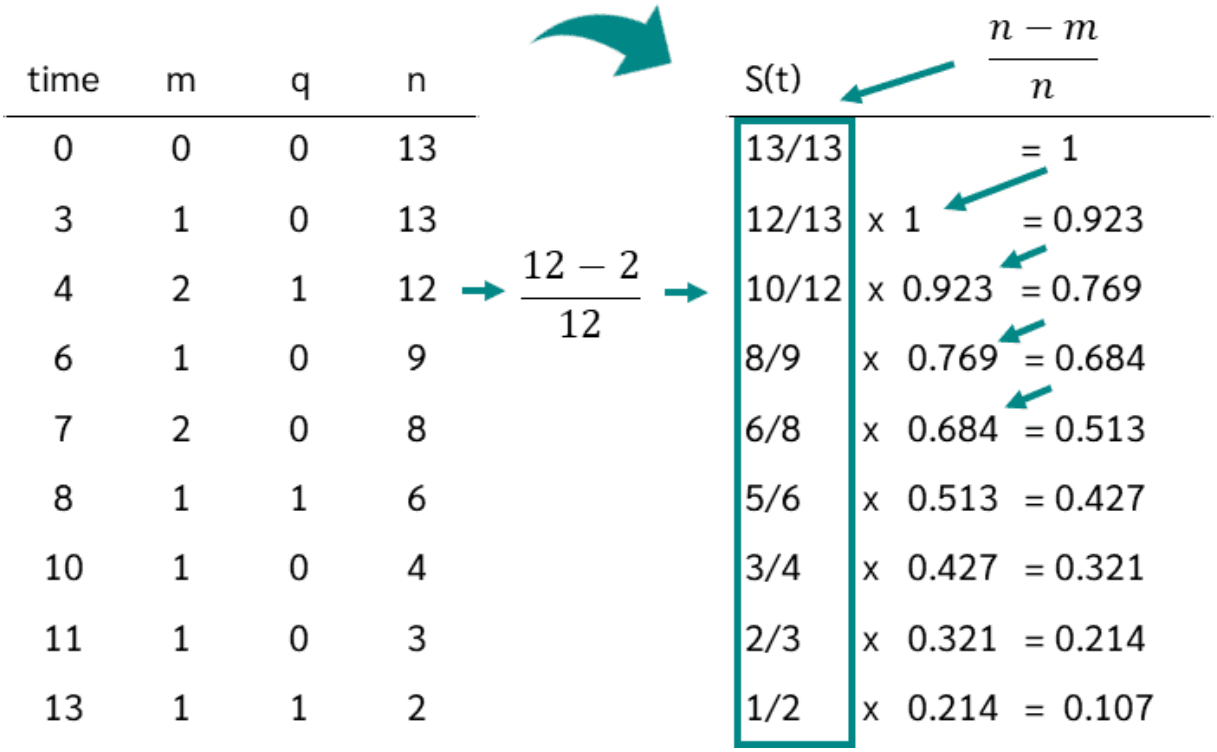
time	Censored
3	1
4	1
4	1
4	0
6	1
7	1
7	1
8	1
9	0
10	1
11	1
13	1
15	0

time	m	q	n
0	0	0	13
3	1	0	13
4	2	1	12
6	1	0	9
7	2	0	8
8	1	1	6
10	1	0	4
11	1	0	3
13	1	1	2

We now need to enter this data into our Kaplan-Meier curve table. We do this as follows: We create our m exactly as we did before, looking at how many cases failed at each time point.

Now we add a column q, in which we enter how many cases were censored at each time.

Note that the time at which each censored case occurred does not get its own row, but is assigned to the previous time.



Let's look at this case. The censoring took place at time 9. In this table, however, there is no event with nine years and we also don't add it. The person is added at time 8.

We can now re-calculate the values for the survival curve. If we have censored data, this is a little more complex.

For this, we write down the values in the first step. We get these values by calculating n-m/n. In the third row, for example, we get the value 10/12 with 12-2 by 12.

The calculation of the real value is iterative. To do this, we multiply the result from the previous row by the value we have just calculated.

So, in the first row we get 1, now we calculate 12/13 times 1, which is equal to 0.923. In the next row we calculate 10/12 times 0.923 and get a value of 0.769. We take this value again for the next row.

We do this for all the rows. We can then plot the Kaplan-Meier curve with this data in the same way as before.

24.6.6 Comparing different groups

If you are comparing several groups or categories (e.g. treatment groups), the Kaplan-Meier curve consists of several lines, each representing a different group. Each line shows the estimated survival rate for that particular group. To test whether there is a statistically significant difference between the groups, the log-rank test can be used.

If you have several factors and you want to see if they have an effect on the curve, you can calculate a Log Rank Test or calculate a Cox Regression here on DATAtab.

24.6.7 Kaplan-Meier curve assumptions

Random or Non-informative censoring: This assumption states that the occurrence of censoring is unrelated to the likelihood of experiencing the event of interest. In other words, censoring should be random and not influenced by factors that affect the event outcome. If censoring is not non-informative, the estimated survival probabilities may be biased.

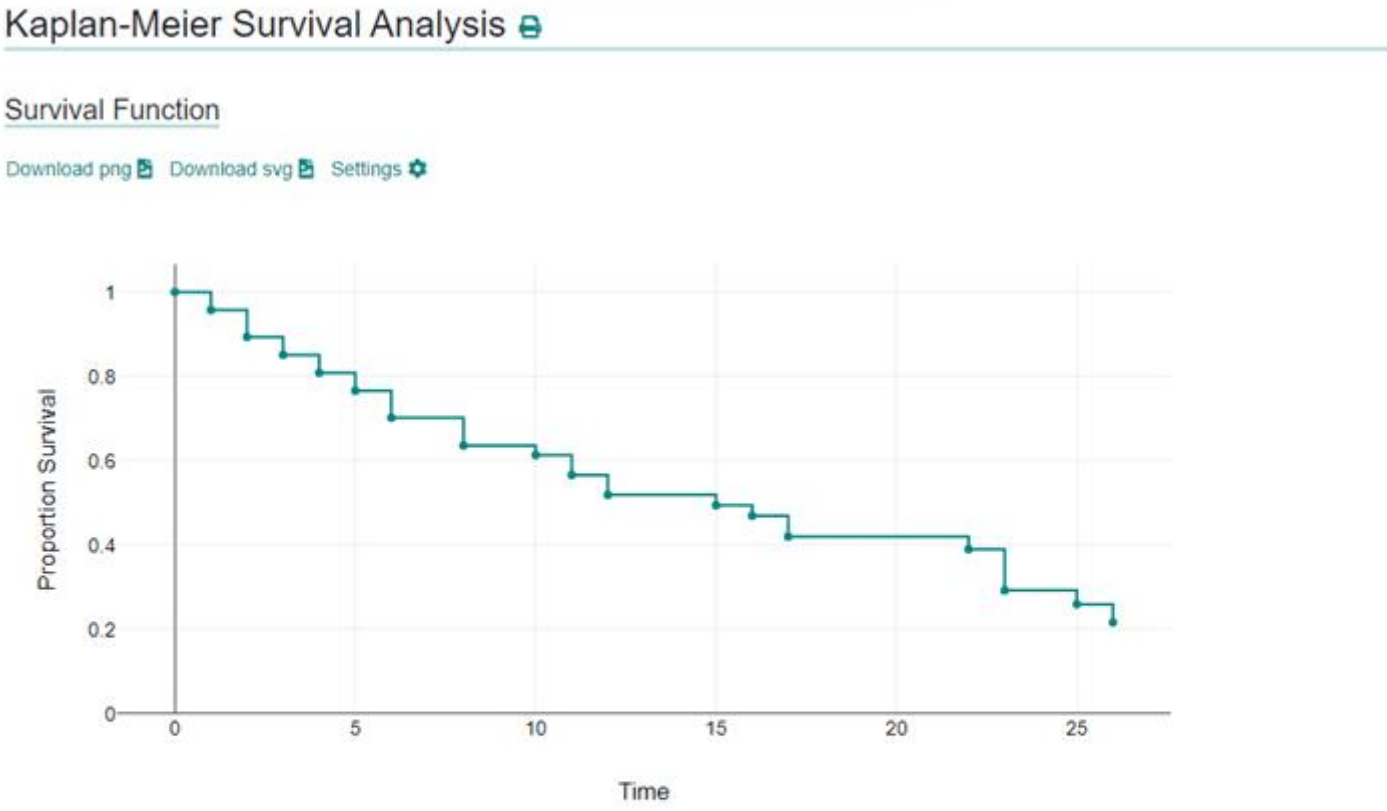
Independence of censoring: This assumption assumes that the censoring times of different individuals are independent of each other. This means that the occurrence or timing of censoring for one participant should not provide any information about the censoring times for other participants.

Survival probabilities do not change over time: The Kaplan-Meier curve assumes that the survival probabilities estimated at each time point remain constant over time. This assumption may not be valid if there are time-varying factors or treatments that can influence survival probabilities.

No competing risks: The Kaplan-Meier curve assumes that the event of interest is the only possible outcome and there are no other competing events that could prevent the occurrence of the event being studied. Competing events can include other causes of death or events that render the occurrence of the event of interest impossible.

24.6.8 Create Kaplan Meier curve with DATAtab

To create the Kaplan Meier curve with DATAtab, simply go to the statistics calculator on datatab.net and copy your own data into the table.



Survival Tabelle

Copy Word Copy Excel

Time	N of Remaining Cases	Event	Censored	Survivor Function	Std. Error	Lower 95% CI	Upper 95% CI
0	47	0	0	1	0	1	1
1	47	2	0	0.96	0.03	0.9	1
2	45	3	0	0.89	0.04	0.81	0.98
3	42	2	0	0.85	0.05	0.75	0.95
4	40	2	0	0.81	0.06	0.7	0.92

Now click on "Plus" and select Survival Analysis. Here you can create the Kaplan Meier curve online. If you select the variable "Time" DATAtab will create the Kaplan Meier curve and you will get the survival table. If you do not click on a status, Datatab assumes that the data is not censored. If this is not the case, click also on the variable that contains the information which case is censored and which is not. One stands for event occurred and 0 stands for censored. Now you will get the appropriate results.

24.7 Log Rank Test

What is the Log Rank Test? The Log Rank Test is used in survival time analysis and compares the distribution of time to event occurrence of two or more independent samples.

With the Log Rank Test you can check if there is a difference between two or more different groups.

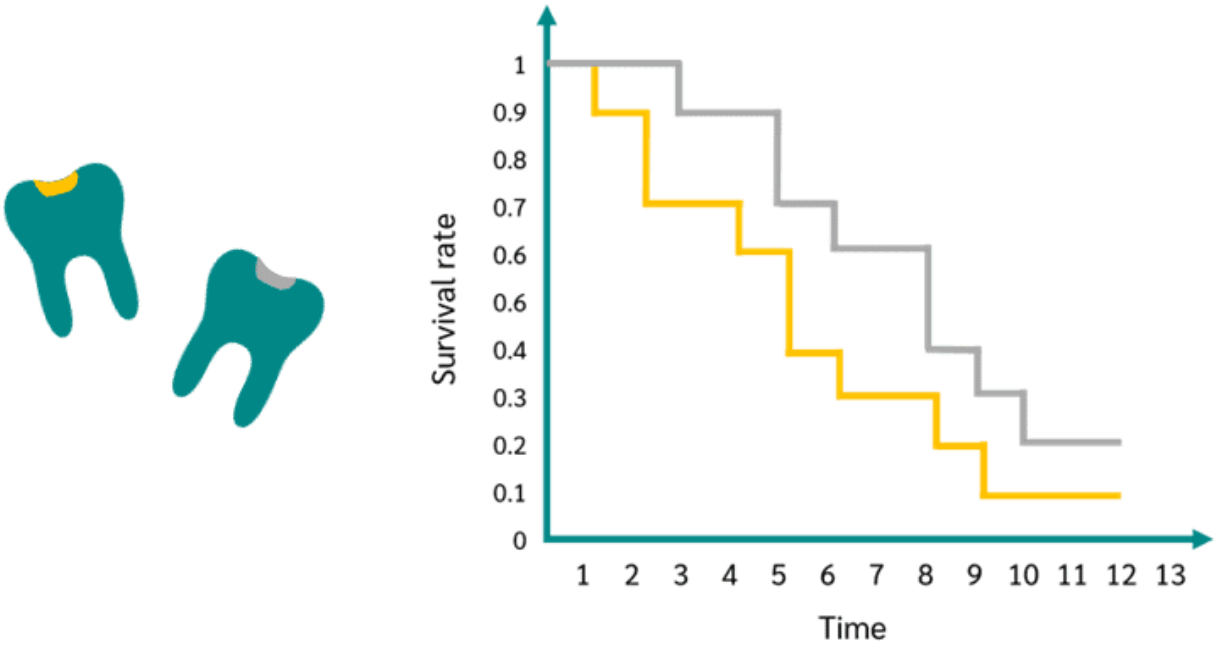


Figure 136: Log Rank Test

What does "distribution" mean? What does "time to event" mean, and what is meant by two or more independent samples? Let's start with the last point, two or more independent samples.

For example, you might want to know if there is a difference between two different materials used for a dental filling.

The next question is, what is the difference? The log rank test checks whether there is a difference in the time it takes for an event to occur.

What does "time to event" mean? The log rank test looks at a variable that has a start time and an end time when a certain event occurs.

Therefore, the log rank test takes into account the time between the start time and the event. This can be measured in days, weeks or months.

In our example, we might be interested in whether the material has an effect on the time it takes for the filling to break out again. We have a starting point, which is the time when the filling is placed. We also have an end point or event, which is the time when the filling breaks out again.

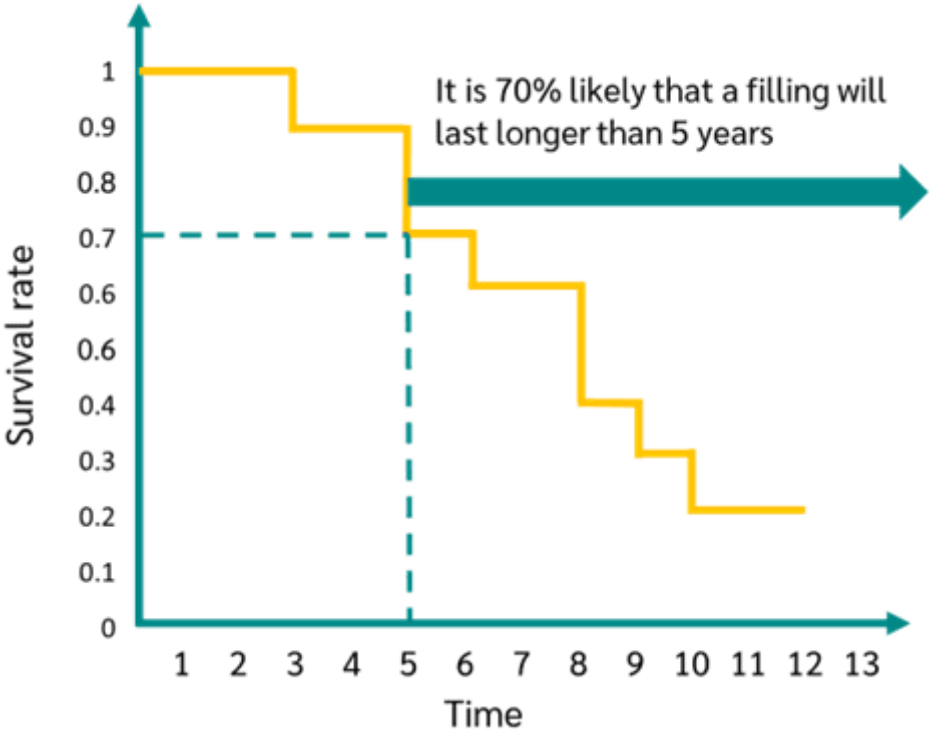
We are interested in the time between the start and the end, that is, the time between the insertion of the filling and the breaking out of the filling.

How do we compare the time it takes for the filling to break out again in each of the test subjects?

We do this using the Kaplan-Meier curve or the table used to create this graph. We plot the time on the x-axis and the survival rate on the y-axis.

What is the survival rate? The Kaplan-Meier curve tells us how likely it is that a filling will last longer than a certain amount of time.

Let's say we want to know how likely it is that a filling will last more than 5 years. In this case, the Kaplan-Meier curve tells you that there is a 70% chance that a restoration will last longer than 5 years.



But now we want to test whether there is a difference between the two materials, so we plot both curves on the graph.

The question that the log rank test answers is: Is there a significant difference between the two curves? In other words, does the filling material have an effect on the "survival time" of the filling?

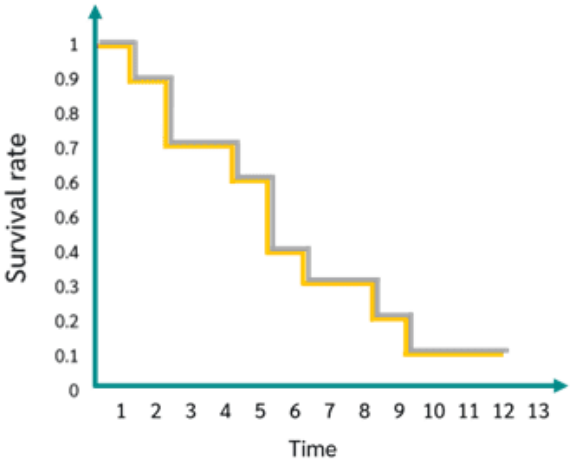
24.7.1 Hypotheses in the Log Rank Test

With this, we can now move on to the null and alternative hypotheses of the log rank test.

- **Null hypothesis:** Both groups have identical distribution curves.
- **Alternative hypothesis:** Both groups have different distribution curves.

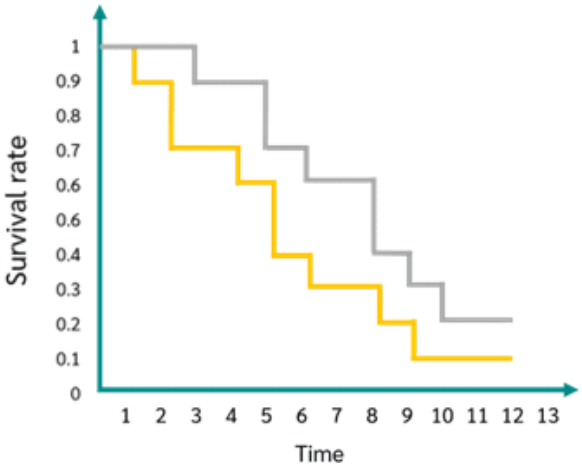
Null hypothesis:

The groups have identical distribution curves.

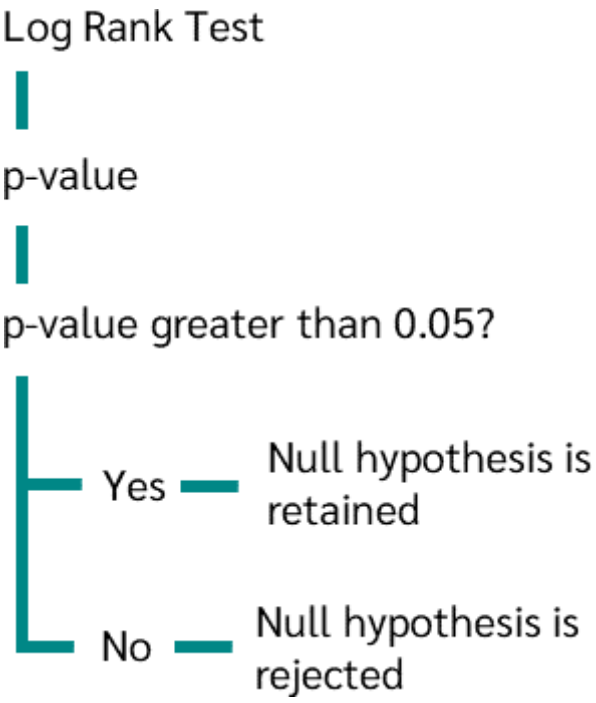


Alternative hypothesis:

The groups have different distribution curves.



So, as always with a statistical hypothesis test, you'll get a p-value out of the log rank test at the end.



The question is whether this p-value is greater than the significance level or not. In most cases, the significance level is set at 0.05.

If the calculated p-value is greater than 0.05, the null hypothesis is retained. Based on the available data, it is then assumed that both groups have the same distribution curve.

If the p-value is less than 0.05, the null hypothesis is rejected and it is assumed that the two groups are different.

24.7.2 Assumptions for the Log Rank Test

The assumptions for the log-rank test are as follows:

Independence: The survival times or event times of individuals in each group should be independent to each other. This assumption implies that the occurrence of an event (e.g., death or failure) for one individual should not influence the occurrence of an event for another individual.

Non-Informative Censoring: Censoring should not be related to the event being studied or to the group assignment (Censored and non-censored patients do not differ in terms of their actual event times). The log-rank test assumes that the probability of censoring should be the same for all individuals within each group. In other words, censoring should not be related to the event being studied or to the group assignment.

Proportional Hazards: The hazard rates (the risk of an event occurring) for the compared groups should be consistent over time. The ratio of the hazard rates should remain constant, indicating that the groups are not experiencing significantly different risks at different time points.

24.7.3 Calculate Log Rank Test

In the next step, we will now discuss the formulas of the Log Rank test and how it is calculated manually.

Suppose we have group 1 and group 2 and we want to test whether both groups have the same survival time function or not.

Group 1		Group 2	
time	Status	time	Status
2	1	2	1
3	1	2	1
5	0	4	1
7	1	4	1
7	1	6	1
8	1	8	1

	Group 1			Group 2		
time	m_1	q_1	n_1	m_2	q_2	n_2
2	1	0	6	2	0	6
3	1	0	5	0	0	4
4	0	1	4	2	0	4
6	0	0	3	1	0	2
7	2	0	3	0	0	1
8	1	0	1	1	0	1

The table above shows the times when either an event occurred or the case was censored. In this case '1' means event occurred '0' means censored.

If we look at our previous example with the fill materials, then each group would have received a different material for the fill. If we assume that time is measured in years, then for group one the first fill would have failed after 2 years, the second fill after 3 years and so on and so forth.

To calculate a log-rank test, we need to combine the tables of Group 1 and Group 2. To do this, we first write down all the time points that appear in the groups.

These are 2, 3, 4, 6, 7 and 8. It is important that the times when only cases were censored are not included in the table. At time 5 a case was censored, but otherwise 5 does not occur, so we do not include time 5 in this table.

Similar to the Kaplan Meier curve, we then fill in the columns m, q and n for groups 1 and 2, respectively. m tells us exactly how many people had an event at that time.

In group 1, one filling broke out after 2 years, one filling broke out after 3 years, nothing happened at time points 4 and 6, two fillings broke out at time point 7, and one filling broke out at time point 8.

q tells us at what time how many cases were censored. Here we only have time 5. As we have already said, we have not entered this time in the table, so this value is assigned to the next earliest time, which is 4, so we have a 1 in the third row. We can do the same for the second group.

From the generated tables we can calculate the so-called expected values for each row. For Group 1 and Group 2 this is done using the following equations.

$$e_1 = \left(\frac{n_1}{n_1 + n_2}\right) \cdot (m_1 + m_2)$$

$$e_2 = \left(\frac{n_2}{n_1 + n_2}\right) \cdot (m_1 + m_2)$$

Let's take a closer look at the first row. *n1* is 6 and *n2* is also 6, so we have 6 divided by 6 plus 6 and *m1* is 1 and *m1* is 2, so we have 1 plus 2. This results in 1.5. We repeat this for all the rows and for both groups.

$$\left(\frac{6}{6+6}\right) \cdot (1+2)$$

m_1	q_1	n_1	m_2	q_2	n_2	e_1	e_2
1	0	6	2	0	6	1.5	1.5
1	0	5	0	0	4	0.56	0.44
0	1	4	2	0	4	1	1
0	0	3	1	0	2	0.6	0.4
2	0	3	0	0	1	1.5	0.5
1	0	1	1	0	1	1	1

Now we need the observed values minus the expected values. For this we simply calculate m1 minus e1 or m2 minus e2.

						Expected		Observed - Expected	
m_1	q_1	n_1	m_2	q_2	n_2	e_1	e_2	$m_1 - e_1$	$m_2 - e_2$
1	0	6	2	0	6	1.5	1.5	-0.5	0.5
1	0	5	0	0	4	0.56	0.44	0.44	-0.44
0	1	4	2	0	4	1	1	-1	1
0	0	3	1	0	2	0.6	0.4	-0.6	0.6
2	0	3	0	0	1	1.5	0.5	0.5	-0.5
1	0	1	1	0	1	1	1	0	0
								Σ	1.15

$$Log\ Rank\ stats = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}$$

Now we can calculate what is called the log rank statistic. We can use either the values from group 1 or the values from group 2. We just take the values from group 2.

O2 minus E2 is obtained by adding these values in the column "m2-e2", which is 1.15. But what is the variance? The variance is given by this formula.

$$Var(O_2 - E_2) = \sum \frac{n_i n_2 (m_1 + m_2)(n_1 + n_2 - m_1 - m_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

$$= 1{,}78$$

We first calculate the following expression for each row and then add them up. In our case we get 1.78.

m_1	q_1	n_1	m_2	q_2	n_2	
1	0	6	2	0	6	0.61
1	0	5	0	0	4	0.25
0	1	4	2	0	4	0.43
0	0	3	1	0	2	0.24
2	0	3	0	0	1	0.25
1	0	1	1	0	1	0
						Σ 1.78

$$\frac{n_i n_2 (m_1 + m_2)(n_1 + n_2 - m_1 - m_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

We can now calculate the log rank statistic. In our example we get 0.74.

$$Log\ Rank\ stats = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}$$

$$= \frac{1{,}15^2}{1{,}78} = \frac{1{,}32}{1{,}78} = 0{,}74$$

The log rank statistic corresponds to a Chi2 value. Therefore, the critical p-value can be determined using the Chi2 distribution. The required degrees of freedom result from the number of groups minus 1.

Log Rank stats

↓

Corresponds to Chi^2 value

↓

p-value is calculated via the Chi^2 distribution

↓

Degrees of freedom result with the number of groups minus 1

24.7.4 Calculate Log Rank Test with DATAtab

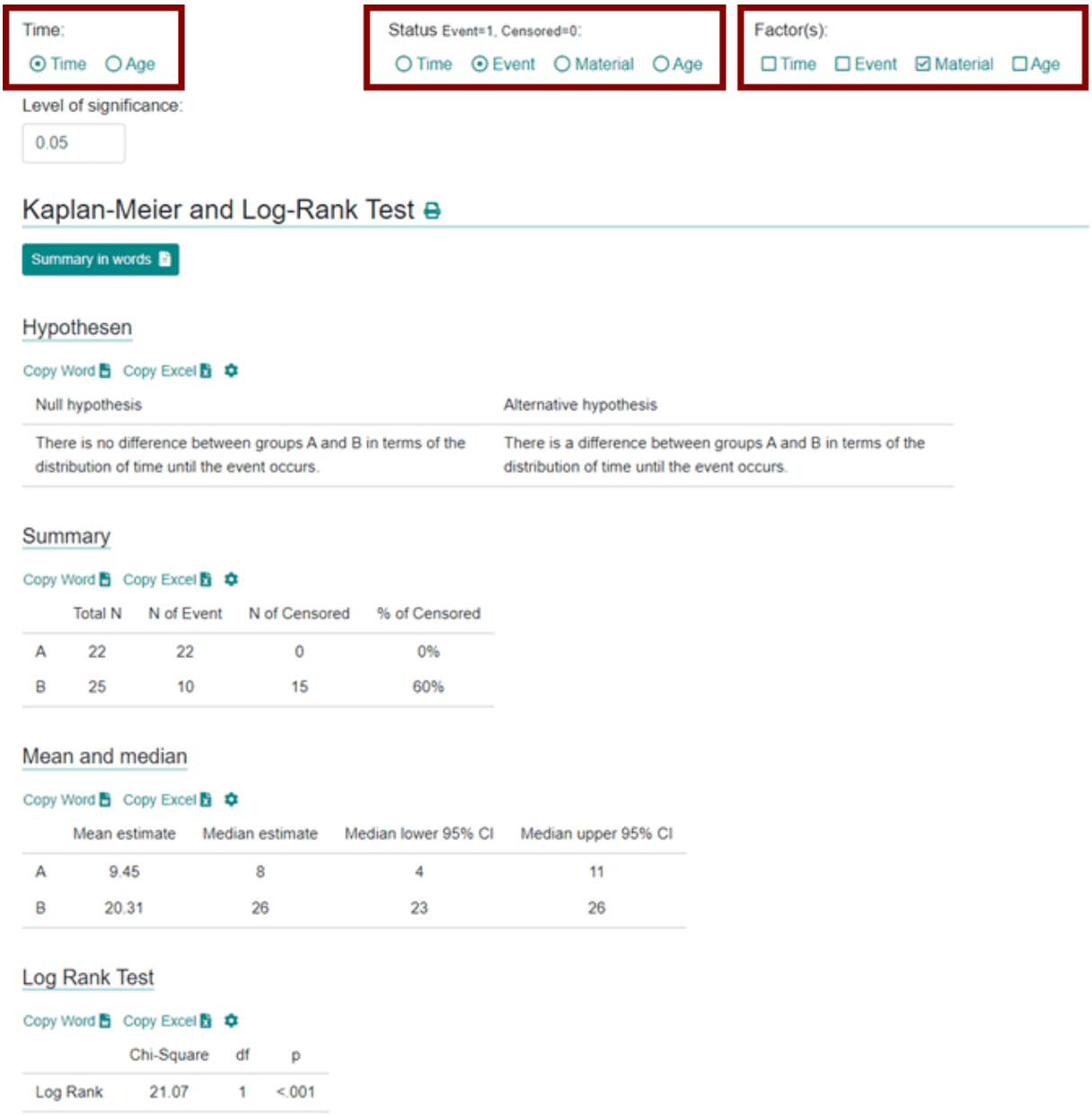
Now you are wondering what is the easiest way to calculate the Log Rank Test? This is best done online with DATAtab. The steps are:

- first you go to the statistics calculator on datatab.net
- copy your own data into the table
- click on "Plus" and click on the tab Survival Analysis

Here we have a column with the time, then a column telling us whether the event occurred or not. Here 1 stands for "occurred" and 0 for "censored". Then we have the variable "Material" with the two materials A and B.

Depending on what you select here, the appropriate methods will be calculated for you. If you select only the variable "Time", the Kaplan-Meier Survival Curve will be displayed with the corresponding table. If you do not select a variable with the status, it is assumed that no case is censored. If this is not the case, you can simply click here at "Status" on the variable that contains the data whether the event has occurred or not.

If now another factor is selected, e.g. the "Material", the log-rank test will be calculated. You can read the null and the alternative hypothesis and get the results of the Log Rank Test listed.



The null hypothesis is: There is no difference between groups A and B in terms of the distribution of time until the event occurs.

And the alternative hypothesis is: There is a difference between groups A and B in the distribution of the time until the event occurs.

Below you can read the results and you can see the p-value for the log rank test. If you don't know exactly how this is interpreted, you can simply click on Summary in words:

A log-rank test was calculated to find out if there is a difference between groups A and B in terms of the distribution of time until the event occurs.

For the data at hand, the log-rank test showed that there is a difference between the groups in terms of the distribution from the time until the event occurs, p=<0.001. The null hypothesis is thus rejected.

This means that if the p-value is greater than the pre-determined significance level, which in most cases is 5%, the null hypothesis is not rejected, i.e. there is then no significant difference.

If the p-value is smaller, the null hypothesis is rejected, and it is assumed on the basis of the available data that there is a difference between the curves.

25. Cox Regression

What is Cox Proportional Hazards Survival Regression or Cox Regression for short? Cox regression is used in survival time analysis to determine the influence of different variables on survival time.

The Variables can be any mixture of continuous, binary, or categorical data. The Cox proportional hazards model is then used to determine the effect on survival time.

Cox Regression allows us to determine the effects of multiple independent variables on a time-to-event outcome, either to test hypotheses about the independent variables or to build a predictive model.

25.1 Survival time analysis

What is survival analysis? In survival time analysis, the survival times of test subjects are recorded and a survival curve is generated. Usually, the subjects have a particular disease.

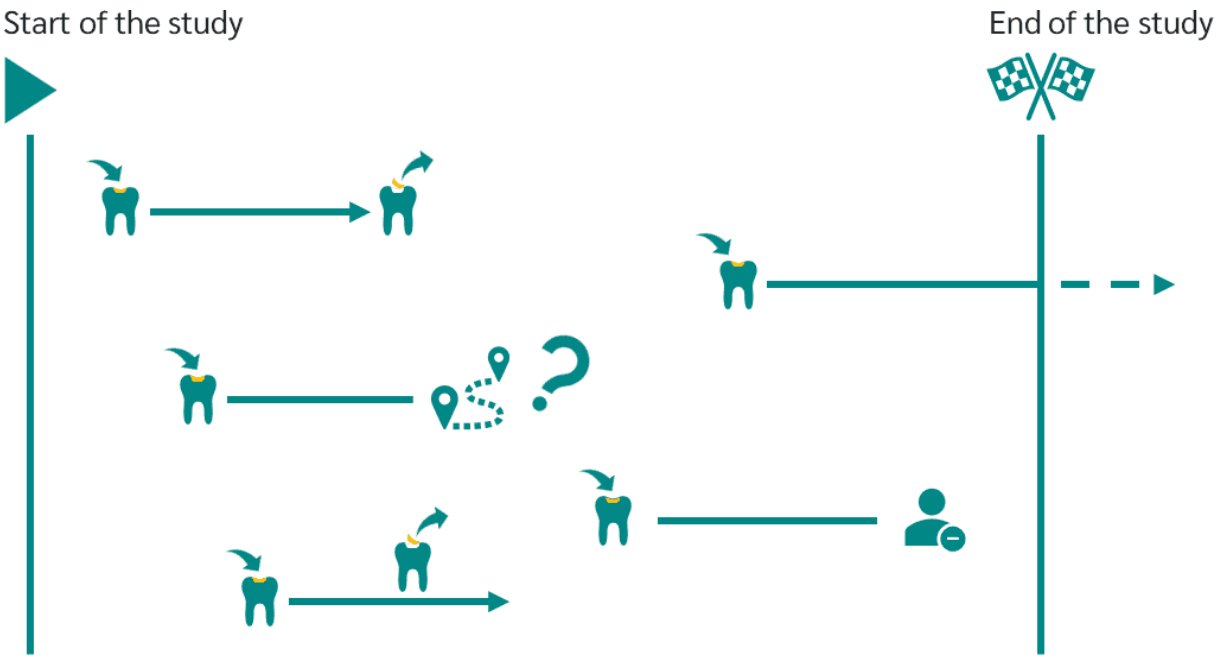
The survival curve then shows how many of the subjects remain alive over time. The considered time does not have to have anything to do with the actual "survival time", nevertheless one speaks of the Survival Time and Survival Time Analysis.

Therefore the survival time analysis considers a variable that has a start time and an end time when a certain event occurs.

The time between the start time and the event is considered in the survival time analysis. This can be measured in days, weeks or months, for example.

25.1.1 Censoring

There is now the problem that a study cannot last indefinitely. This results from limited time and financial resources and from the fact that one would like to publish the results at some point. Therefore, each study has a start date and an end date. If there is no clear event date for a case, it is referred to as "censoring".



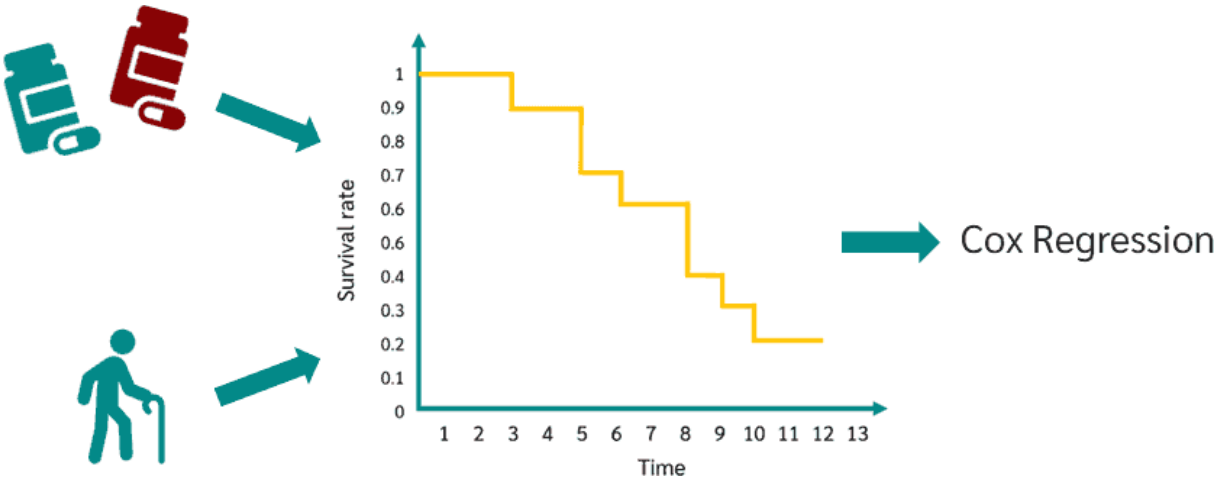
25.1.2 Cox Regression Example

Let's go back to the Cox regression. For example, if you want to analyze the survival time after the detection of a disease, you are often not interested in the survival time itself, but in what influences the survival time.

So, we want to know if the survival time depends on one or more factors, called "predictors" or "independent variables".

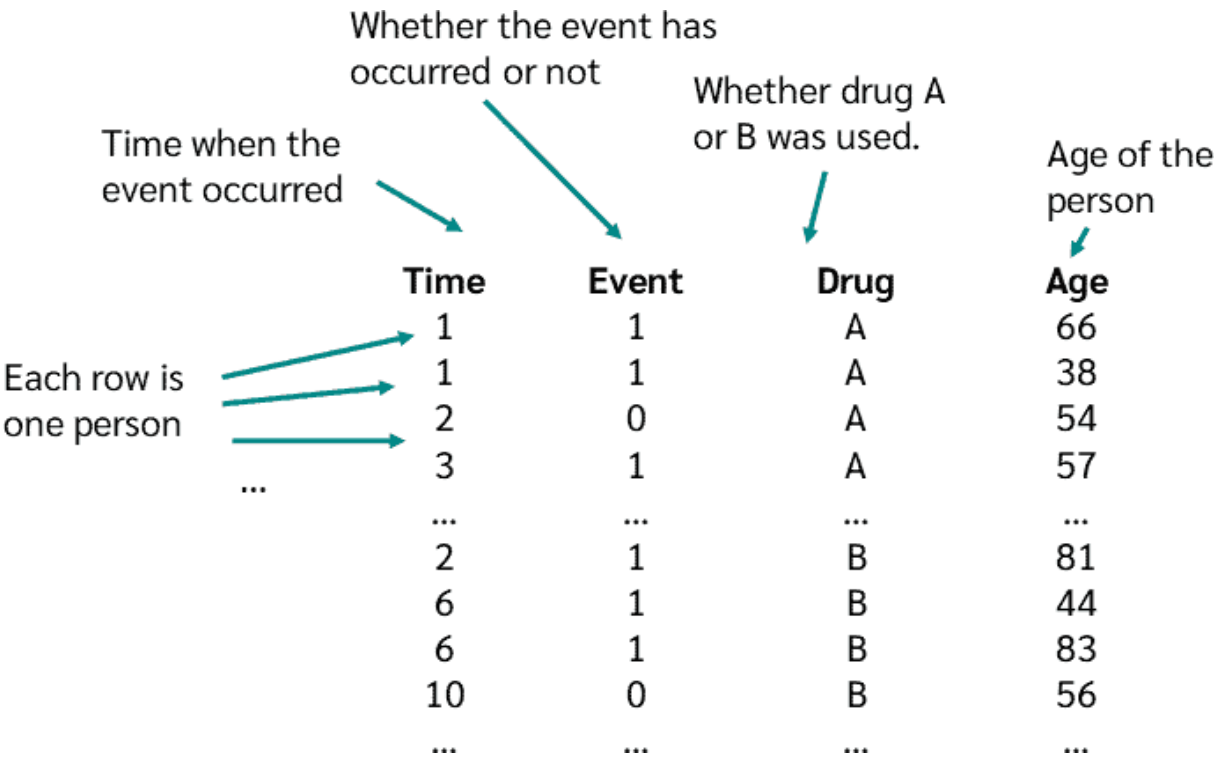
For simple situations with a single factor with only two values, the Log Rank Test is used. For example, if you want to test whether there is a difference in survival time when two different drugs are given.

If you want to include the age of the subjects, a special type of regression is needed. This is the Proportional Hazards Survival Regression. This regression is then used to evaluate the effect of each predictor on the shape of the survival curve.



In our example, we have as predictors on the one hand the drug used and on the other hand the age of the persons. We would now like to know what influence these variables have on the survival time curve. For this purpose, we resort to Cox regression.

We will now take a look at the individual steps of the Cox regression using an example. Let's assume that we have the following data and we want to evaluate them.



Each row describes a patient with the corresponding disease. The time indicates when the event or death occurred. Of course, we also have the information about which drug was used and the age of the subjects.

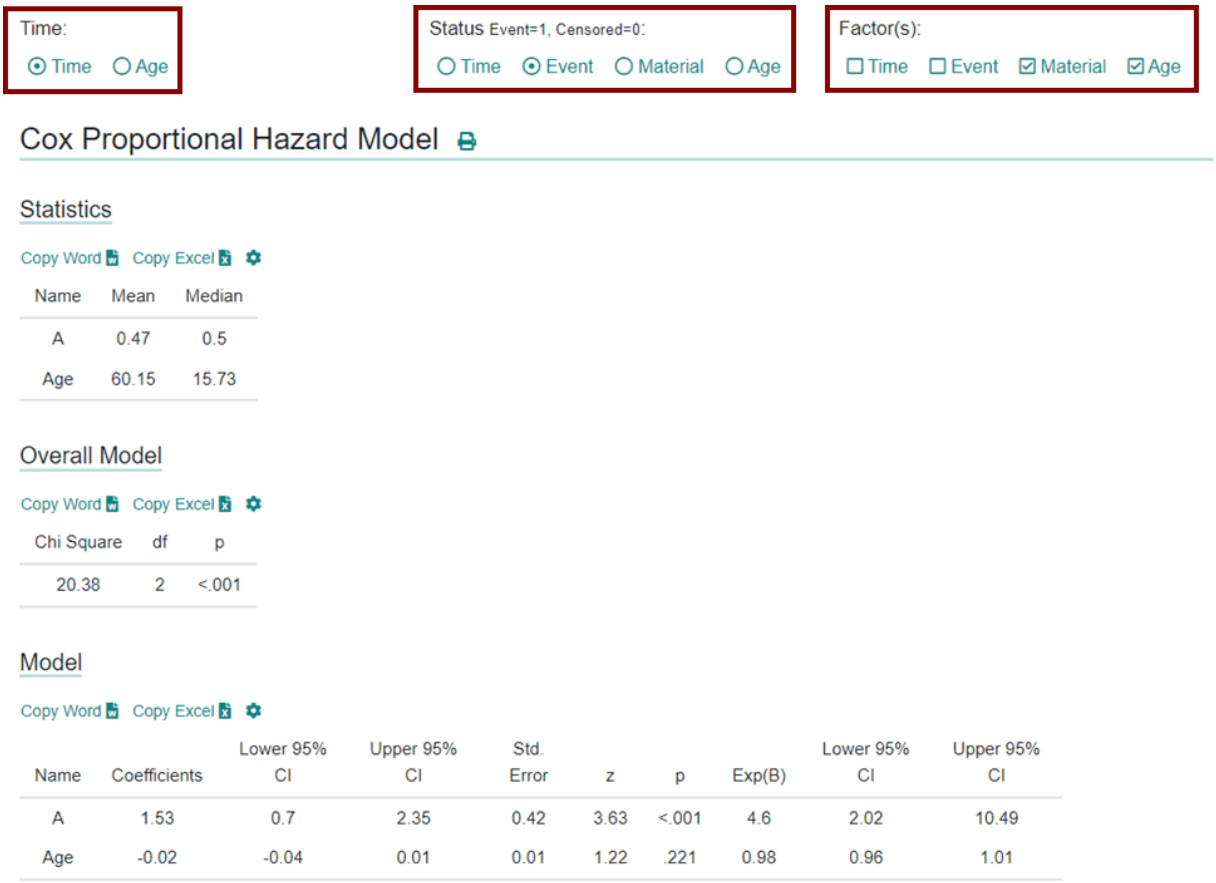
25.1.3 Calculate Cox Regression with DATAtab

The first step is to calculate the Cox regression, we will do this online using DATAtab, then we will go through how to interpret the results. Please load the data above.

To calculate the Cox Proportional Hazards Survival Regression with your own data, simply go to the Cox Regression Calculator and copy and paste your data into the table as you would in Excel.

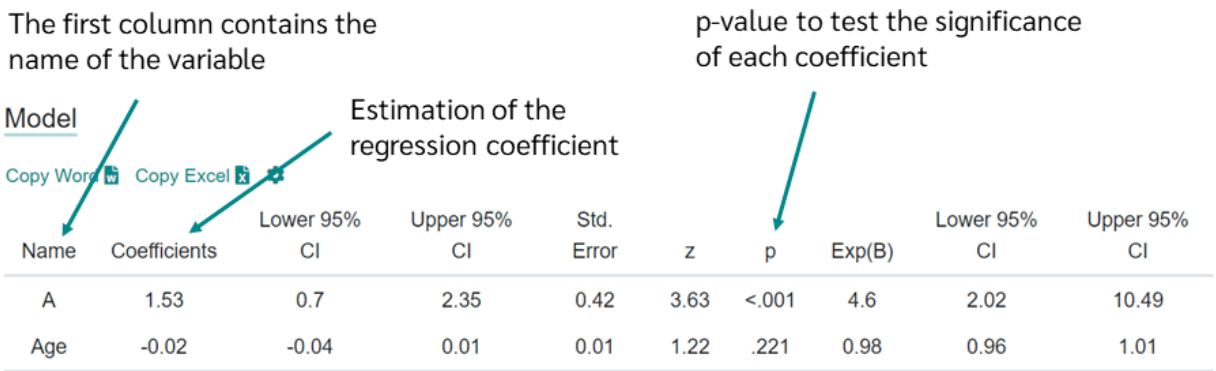
Now we click on "Survival Analysis." Depending on which variables you want to select, different methods of survival analysis will be calculated. If you select only the "Time" and the "Status", the Kaplan Meier curve will be displayed.

If you now click on the drug, you will get the log rank test. If you also select the age, the Cox regression will be calculated.



25.1.4 Interpretation of the Cox Regression

Let's look at the results of a Cox regression. The first column shows the names of the variables. The first row shows the variable drug and the second row shows the age of the persons.


The first column contains the name of the variable
p-value to test the significance of each coefficient
Model
Estimation of the regression coefficient
Copy Word
Copy Excel
Name Coefficients Lower 95% CI Upper 95% CI Std. Error z p Exp(B) Lower 95% CI Upper 95% CI
A 1.53 0.7 2.35 0.42 3.63 <.001 4.6 2.02 10.49
Age -0.02 -0.04 0.01 0.01 1.22 .221 0.98 0.96 1.01

The most important values in this table are the estimated regression coefficient and the p-value. With the help of the p-value you can read whether the regression coefficient is significantly different from zero.

The null hypothesis is therefore that in the population the coefficient is zero. Assuming, as usual, that the significance level is set at 5%, then the null hypothesis is rejected for p-values less than 5 or 0.05. The coefficient is thus significantly different from zero.

In the case of the drug, the p-value is less than 0.05 and thus there is a significant difference of zero.

In the case of age, we obtain a p-value of 0.221, which is thus greater than 0.05. Therefore, in this case, the null hypothesis is not rejected or retained, and we assume from these data that age has no significant effect on the survival curve.

25.1.5 Assumptions of a Cox Regression

Proportional Hazards Assumption: The proportional hazards assumption is the central assumption of Cox regression. It states that the hazard ratio (the ratio of the hazard rates between two groups) remains constant over time. In other words, the effect of the predictor variables on the hazard function is assumed to be constant over time.

Independence Assumption: Cox regression assumes that the survival times of individuals are independent of each other, given the values of the predictor variables. This means that the survival time of one individual should not influence the survival time of another individual.

Linearity Assumption: Cox regression assumes that the relationship between the predictor variables and the log of the hazard rate is linear. This assumption implies that the effect of a continuous predictor is constant over its entire range.

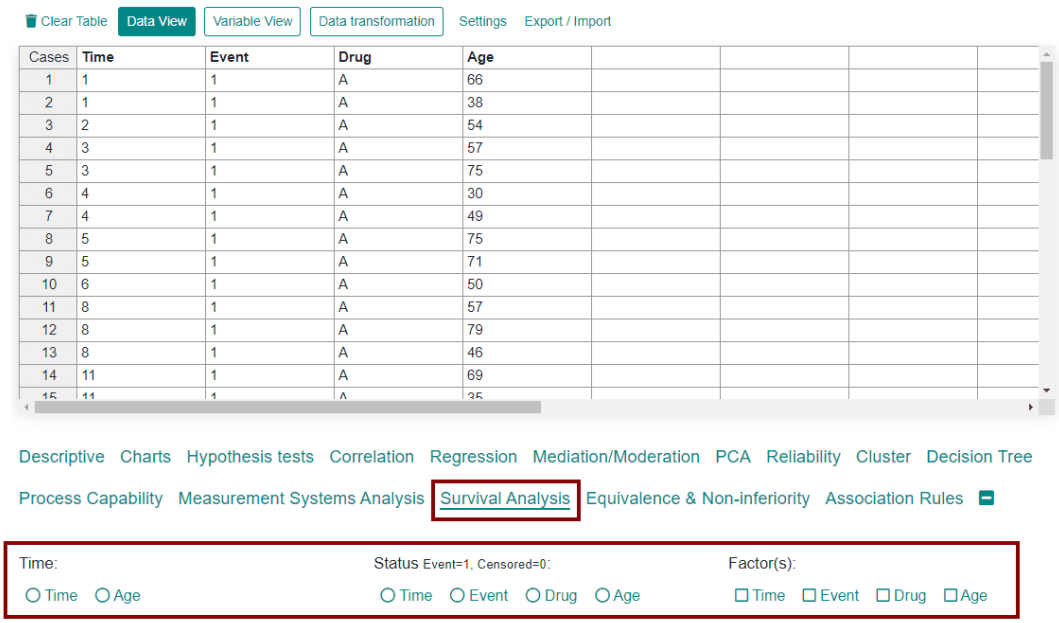
No Multicollinearity: Cox regression assumes that there is no perfect multicollinearity among the predictor variables. Multicollinearity occurs when two or more predictor variables are highly correlated, making it difficult to separate their individual effects on the outcome.

No Outliers: Cox regression assumes that there are no extreme outliers that significantly affect the results. Outliers are observations that deviate substantially from the overall pattern of the data and can distort the estimated coefficients.

No Effect Modification: Cox regression assumes that there is no effect modification or interaction between the predictor variables. Effect modification occurs when the effect of one predictor variable on the outcome depends on the level of another predictor variable.

25.1.6 Calculate survival time analysis with DATAtab

With DATAtab you can easily calculate a survival analysis online. Just go to (1) datatan.de, (2) copy your own data into the table, and (3) click on "Plus" and then on Survival Analysis.



Cases	Time	Event	Drug	Age
1	1	1	A	66
2	1	1	A	38
3	2	1	A	54
4	3	1	A	57
5	3	1	A	75
6	4	1	A	30
7	4	1	A	49
8	5	1	A	75
9	5	1	A	71
10	6	1	A	50
11	8	1	A	57
12	8	1	A	79
13	8	1	A	46
14	11	1	A	69
15	11	1	A	35

In the example above we have a column with the "time", then a column that tells us whether the "event has occurred" or not, i.e. whether the case is censored or not. Here 1 stands for "occurred" and 0 for "censored".

Then we have the variable "Material" with the two materials A and B and we have the "Age". Depending on what you click here, the appropriate methods are calculated.

If you only select the variable "Time", the Kaplan-Meier survival curve will be displayed, and you will get the corresponding survival time table. If no variable is specified with the status, the calculation assumes that no case is censored.

If this is not the case, you can simply click on the variable "Status", which contains the information about whether the event has occurred or not.

Time:
Time Age

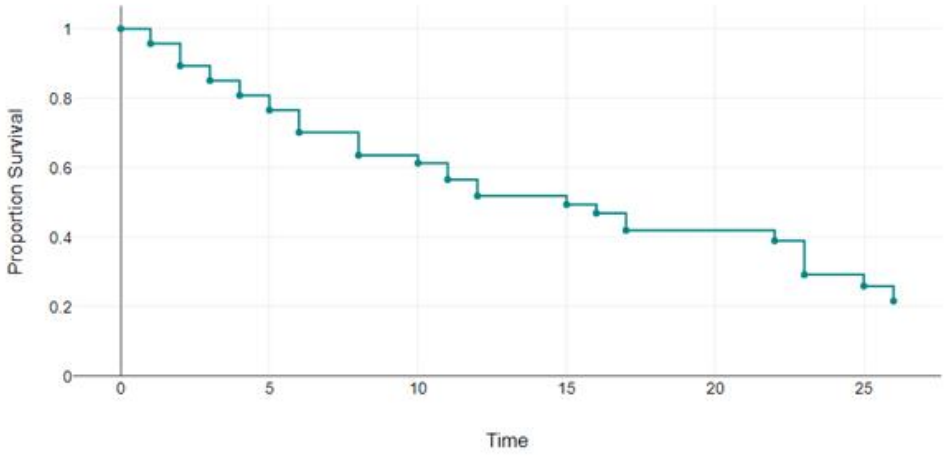
Status Event=1, Censored=0:
Time Event Drug Age

Factor(s):
Time Event Drug Age

Kaplan-Meier Survival Analysis

Survival Function

Download png Download svg Settings

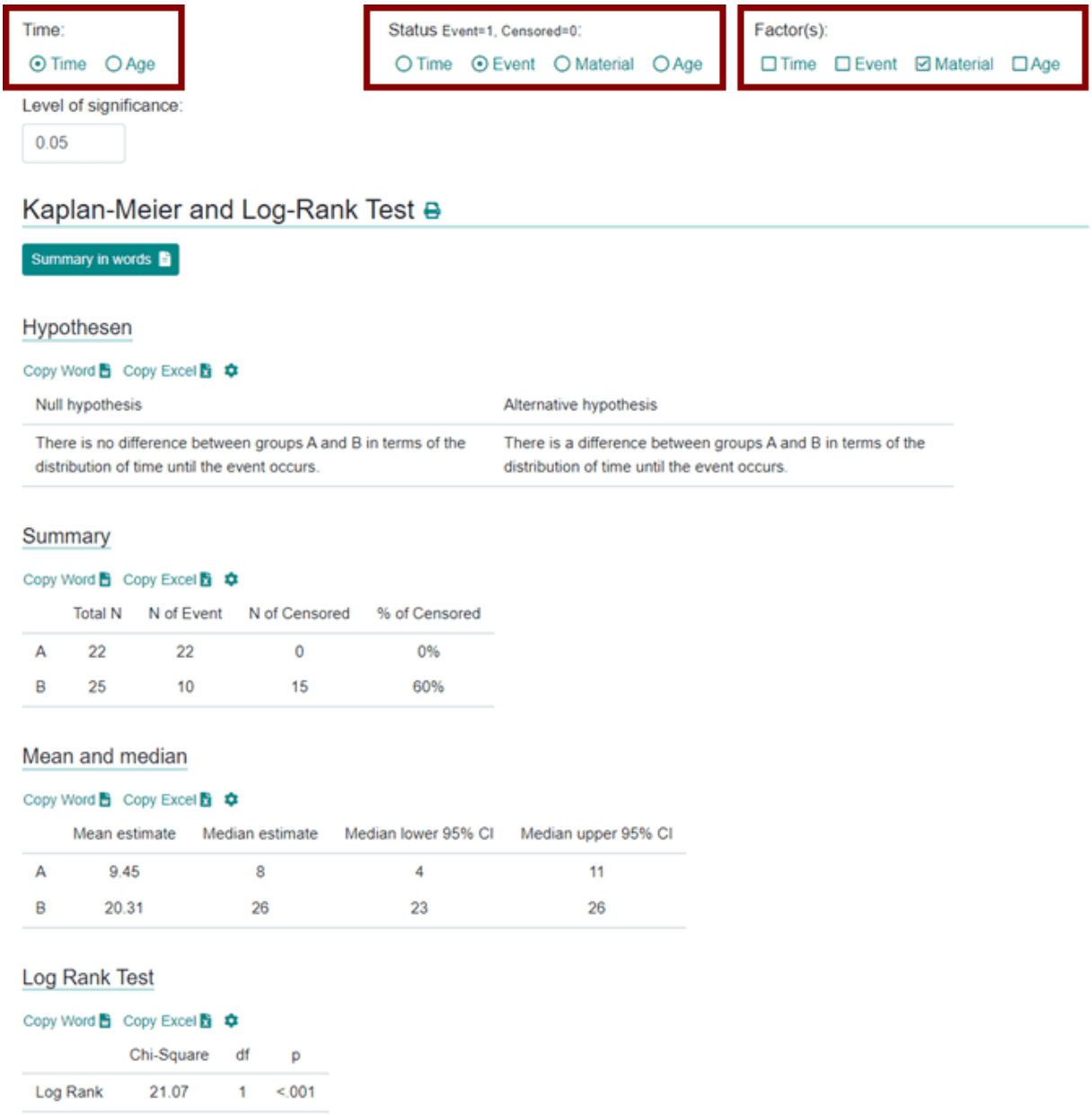


Survival Tabelle

Copy Word Copy Excel

Time	N of Remaining Cases	Event	Censored	Survivor Function	Std. Error	Lower 95% CI	Upper 95% CI
0	47	0	0	1	0	1	1
1	47	2	0	0.96	0.03	0.9	1
2	45	3	0	0.89	0.04	0.81	0.98
3	42	2	0	0.85	0.05	0.75	0.95
4	40	2	0	0.81	0.06	0.7	0.92

If another factor is selected, e.g. the "material", the log-rank test is calculated. Then you get the null and the alternative hypothesis as well as the p-value for the long rank test.



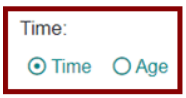
The null hypothesis is as follows: There is no difference between groups A and B in terms of the distribution of time until the event occurs.

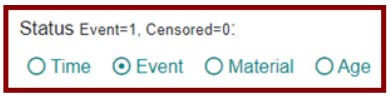
If you go further down in the results section, you will find the p-value. If you don't know exactly how this is interpreted, you can simply click on "Summary in words":

A log-rank test was calculated to see if there was a difference between groups A and B in terms of the distribution of time until the event occurs.

For the present data, the log-rank test revealed that there is a difference between the groups in terms of the distribution of time until the event occurs, p=<0.001. Thus, the null hypothesis is rejected.

On the other hand, if the "material" and the "age" were selected, the Cox regression is calculated. Then you can read whether the factors have a significant influence or not.





Cox Proportional Hazard Model

Statistics

Copy Word Copy Excel

Name	Mean	Median
A	0.47	0.5
Age	60.15	15.73

Overall Model

Copy Word Copy Excel

Chi Square	df	p
20.38	2	<.001

Model

Copy Word Copy Excel

Name	Coefficients	Lower 95% CI	Upper 95% CI	Std. Error	z	p	Exp(B)	Lower 95% CI	Upper 95% CI
A	1.53	0.7	2.35	0.42	3.63	<.001	4.6	2.02	10.49
Age	-0.02	-0.04	0.01	0.01	1.22	.221	0.98	0.96	1.01

26. z-Score

This tutorial is about z-standardization (z-transformation). We discuss what the z-score is, how z-standardization works and what the standard normal distribution is. It also explains what the z-score table is and what it is used for.

26.1 What is z-standardization?

Z-standardization is a statistical procedure used to make data points from different datasets comparable. In this procedure, each data point is converted into a z-score. A z-score indicates how many standard deviations a data point is from the mean of the dataset.

26.2 Example of z-standardization

Suppose you are a doctor and want to examine the blood pressure of your patients. For this purpose, you measured the blood pressure of a sample of 40 patients. From the measured data, you can now naturally calculate the average, i.e., the value that the 40 patients have on average.

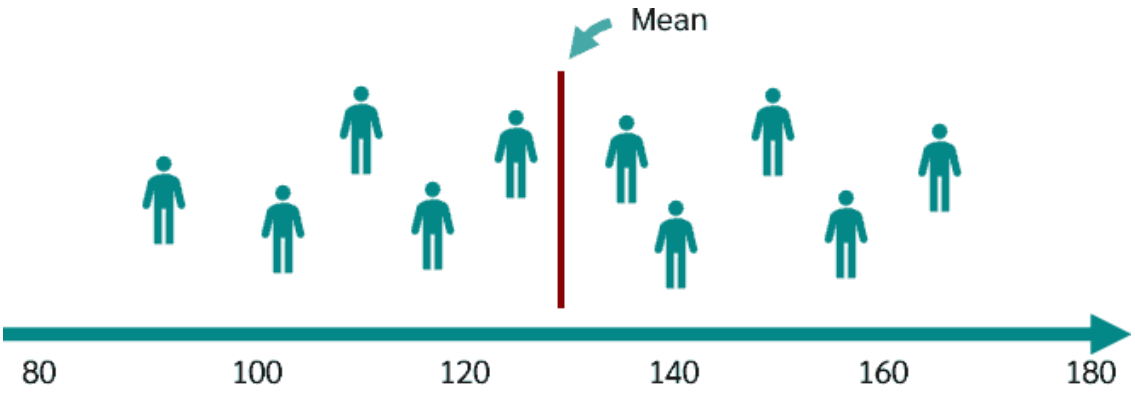
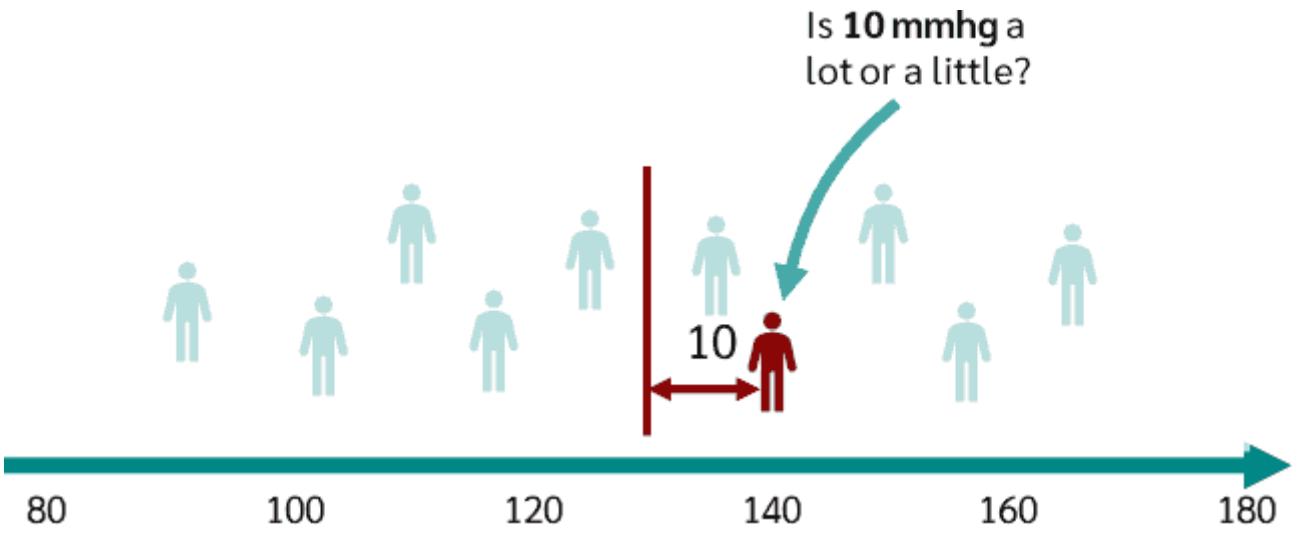
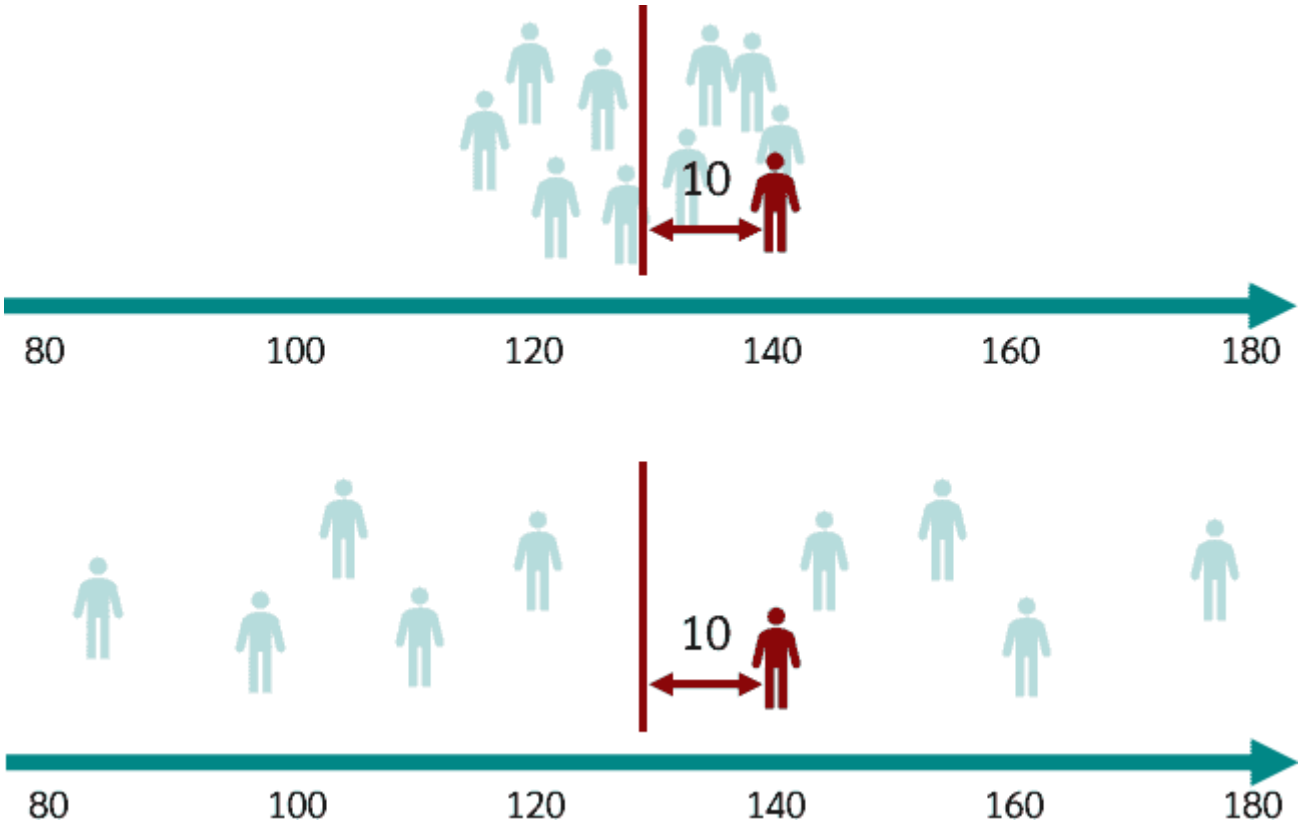


Figure 137: z-standardization example

Now one of the patients asks you how high his blood pressure is compared to the others. You tell him that his blood pressure is 10mmHg above average. Now the question arises, whether 10mmHg is a lot or a little.

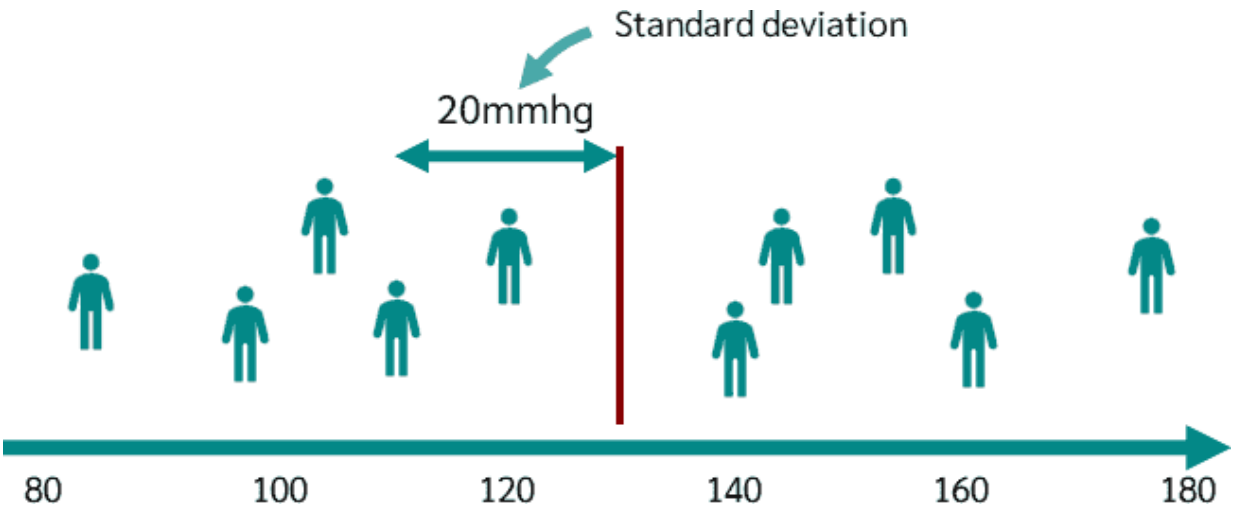


If the other patients cluster very closely around the mean, then 10mmHg is a lot in relation to the spread, but if the other patients spread very widely around the mean, then 10mmHg might not be that much.

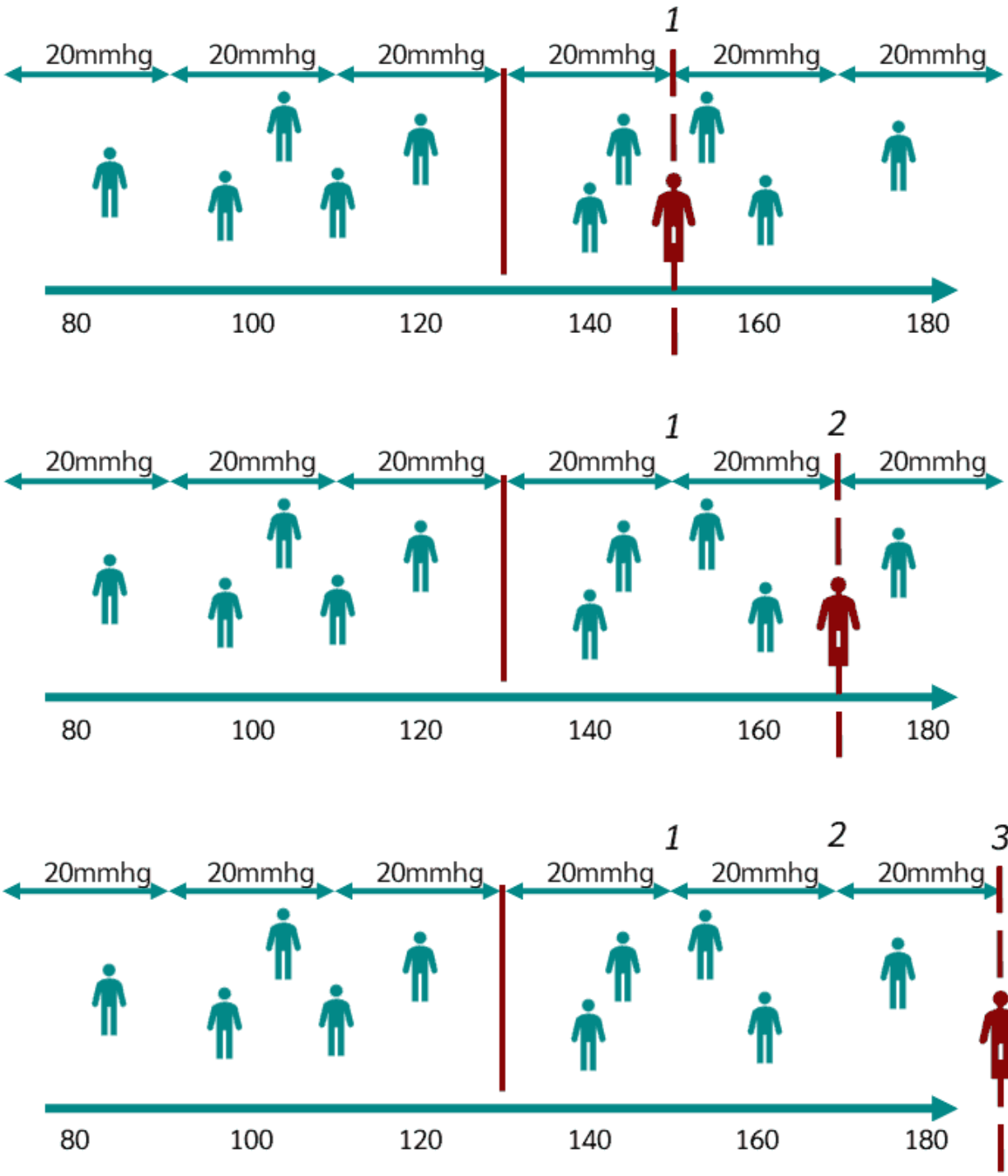


The standard deviation tells us how much the data is spread out. If the data are close to the mean, we have a small standard deviation; if they are widely spread, we have a large standard deviation.

Let's say we get a standard deviation of 20 mmHg for our data. This means that on average, the patients deviate by 20 from the mean.

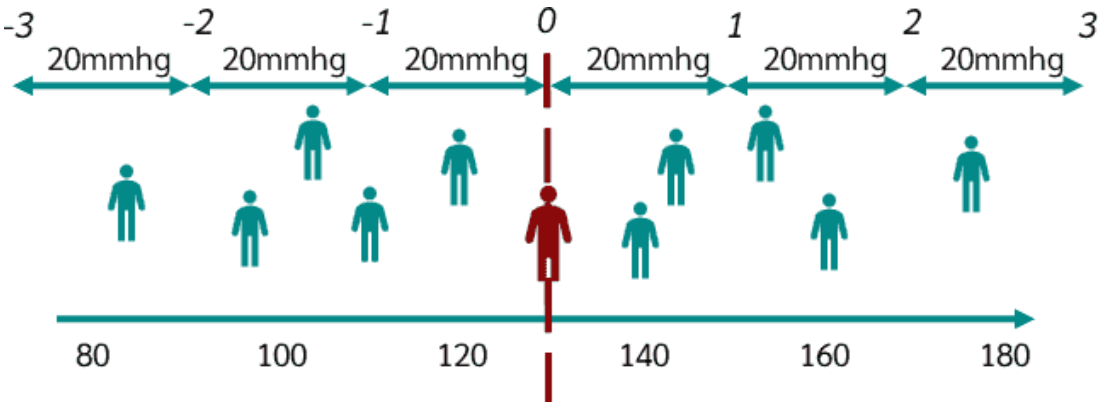


The z-score now tells us how far a person is from the mean in units of standard deviation. So, a person who deviates one standard deviation from the mean has a z-score of 1. A person who is twice as far from the mean has a z-score of 2. And a person who is three standard deviations from the mean has a z-score of 3.


1
20mmhg
20mmhg
20mmhg
20mmhg
20mmhg
20mmhg
80
100
120
140
160
180
1
2
20mmhg
20mmhg
20mmhg
20mmhg
20mmhg
20mmhg
80
100
120
140
160
180
1
2
3
20mmhg
20mmhg
20mmhg
20mmhg
20mmhg
20mmhg
80
100
120
140
160
180

Accordingly, a person who deviates by minus one standard deviation has a z-score of -1, a person who deviates by minus two standard deviations has a z-score of -2, and a person who deviates by minus three standard deviations has a z-score of -3.

And if a person has exactly the value of the mean, then they deviate by zero standard deviations from the mean and receive a score of zero.



Thus, the z-score indicates how many standard deviations a measurement is from the mean. As mentioned, the standard deviation is just a measure of the dispersion of the patients' blood pressure around the mean.

In short, the z-score helps us understand how exceptional or normal a particular measurement is compared to the overall average.

26.3 Calculating the z-score

How do we calculate the z-score? We want to convert the original data, in our case the blood pressure, into z-scores, i.e., perform a z-standardization.

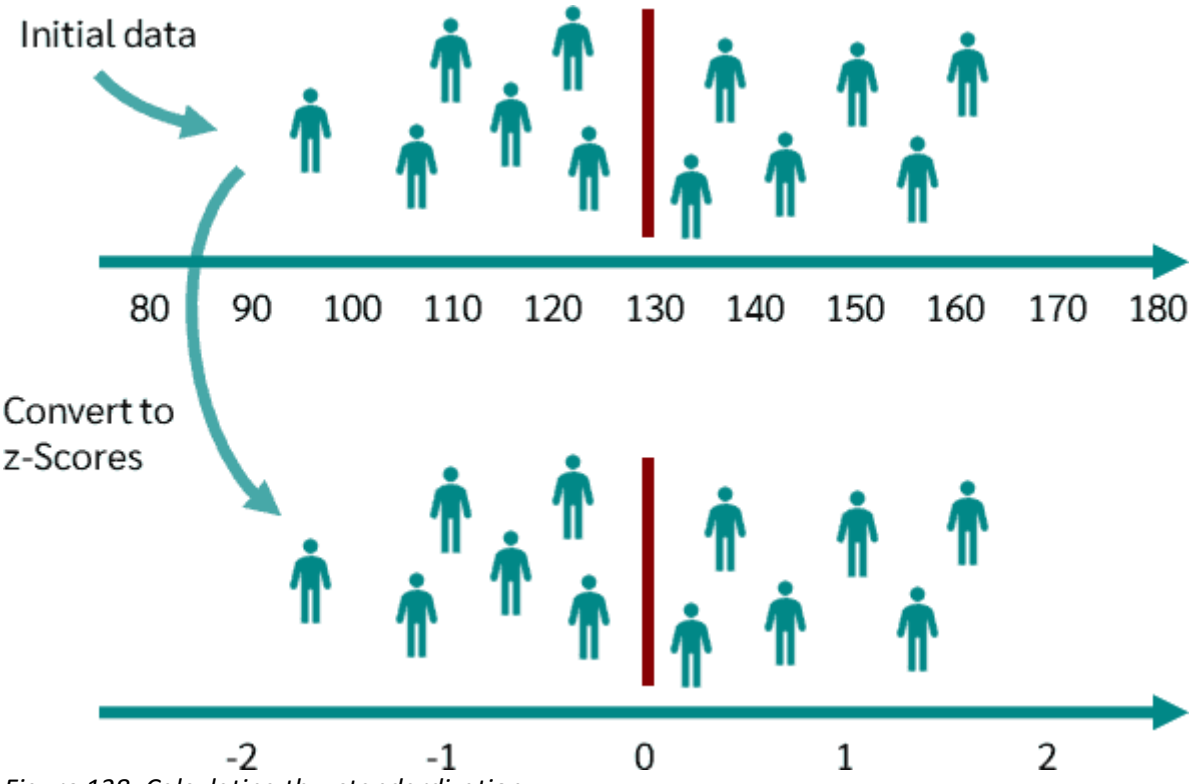
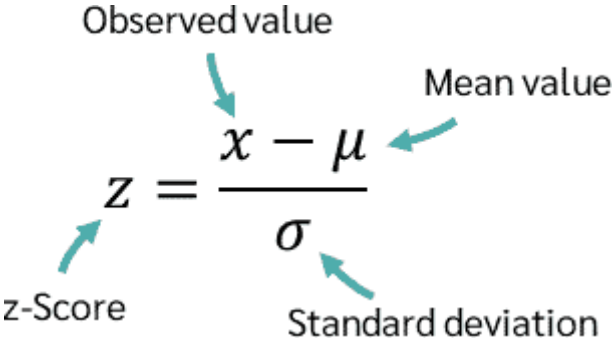


Figure 138: Calculating th z-standardization

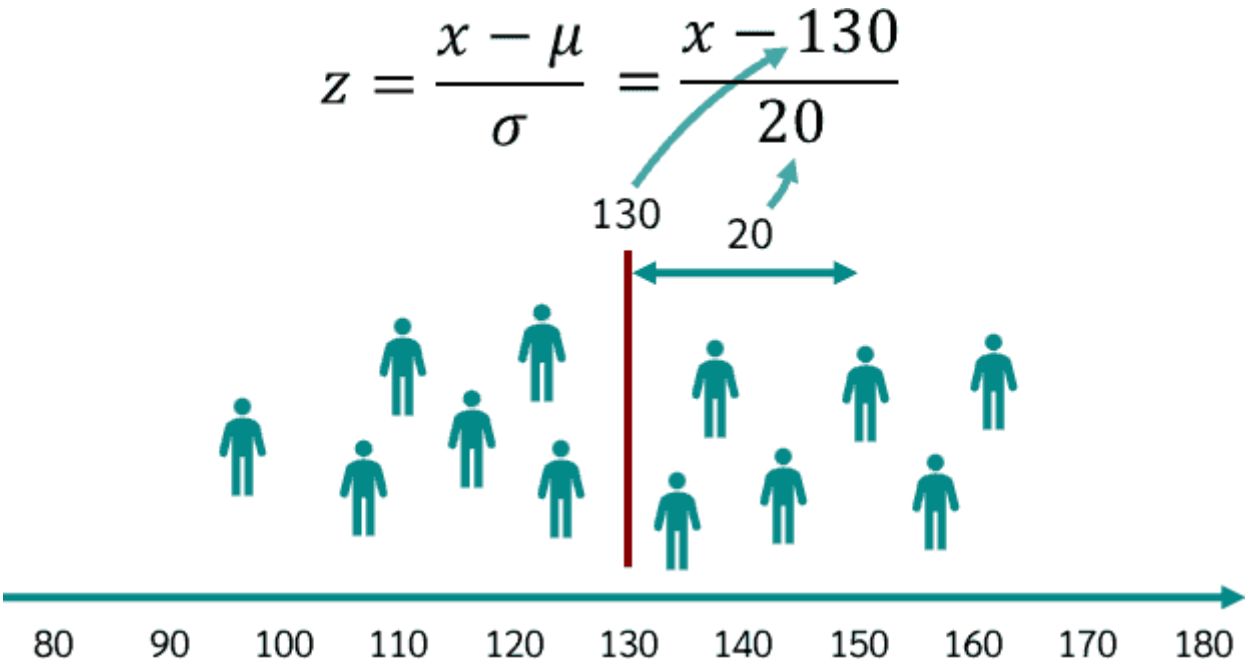
Here we see the formula for z-standardization. Here, z is of course the z-value we want to calculate, x is the observed value, in our case the blood pressure of the person in question, μ is the mean value of the sample, in our case the mean value of all 40 patients, and σ is the standard deviation of the sample, i.e. the standard deviation of our 40 patients.



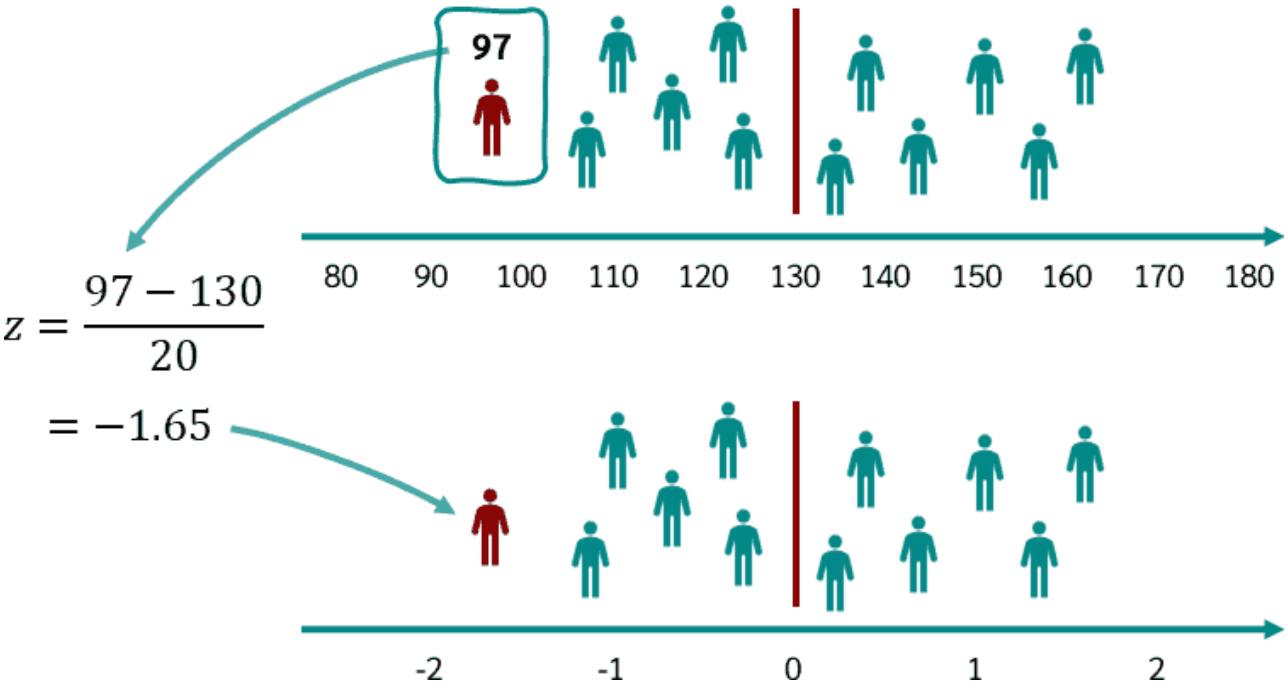
Caution: μ and σ are actually the mean and standard deviation of the population, but in our case we only have a sample. However, under certain conditions, which we will discuss later, we can estimate the mean and standard deviation using the sample.

Let's assume that the 40 patients in our example have a mean value of 130 and a standard deviation of 20. If we use both values, we get for z: x-130 divided by 20

$$z = \frac{x - \mu}{\sigma} = \frac{x - 130}{20}$$



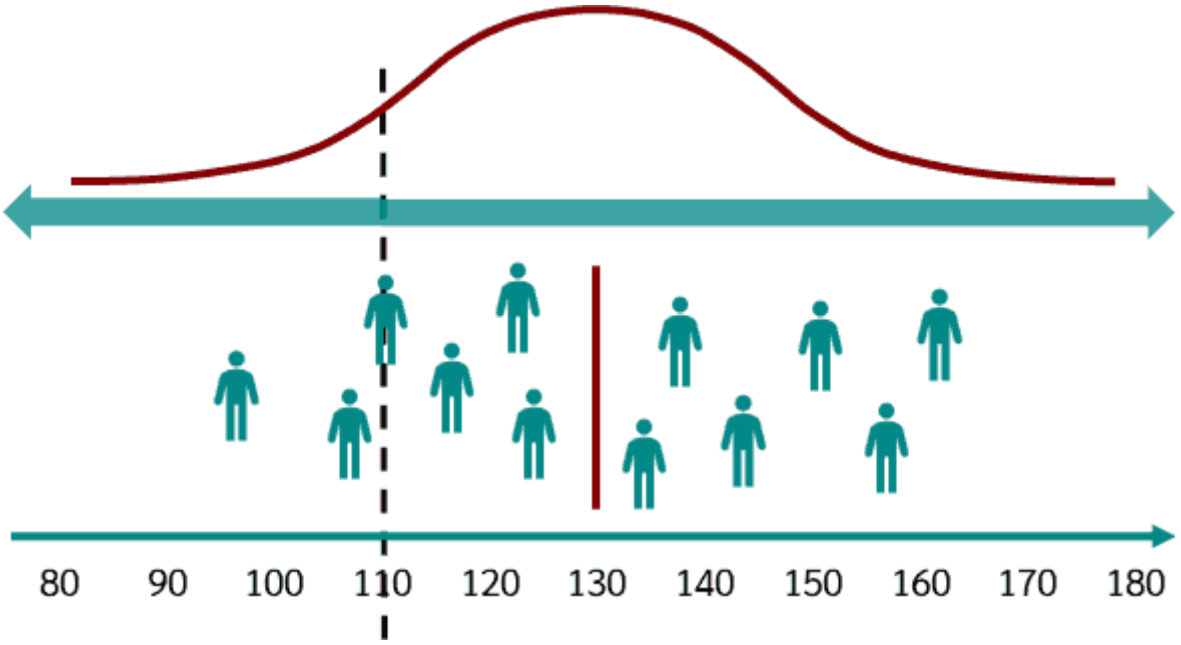
Now we can use the blood pressure of each individual patient for x and calculate the z value. Let's just do this for the first patient. Let's say this patient has a blood pressure of 97, then we simply enter 97 for x and get a z-value of -1.65.



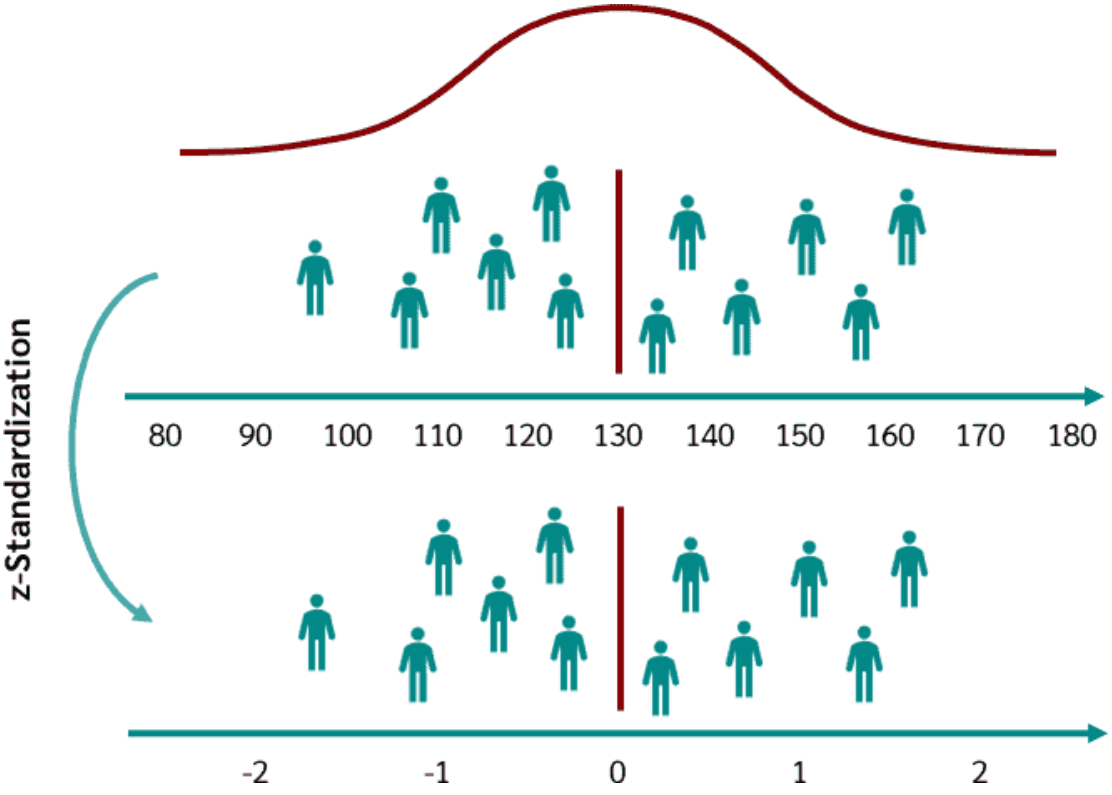
This person therefore deviates from the mean by -1.65 standard deviations. We can now do this for all patients.

Regardless of the unit of the initial data, we now have an overview in which we can see how far a person deviates from the mean in units of the standard deviation.

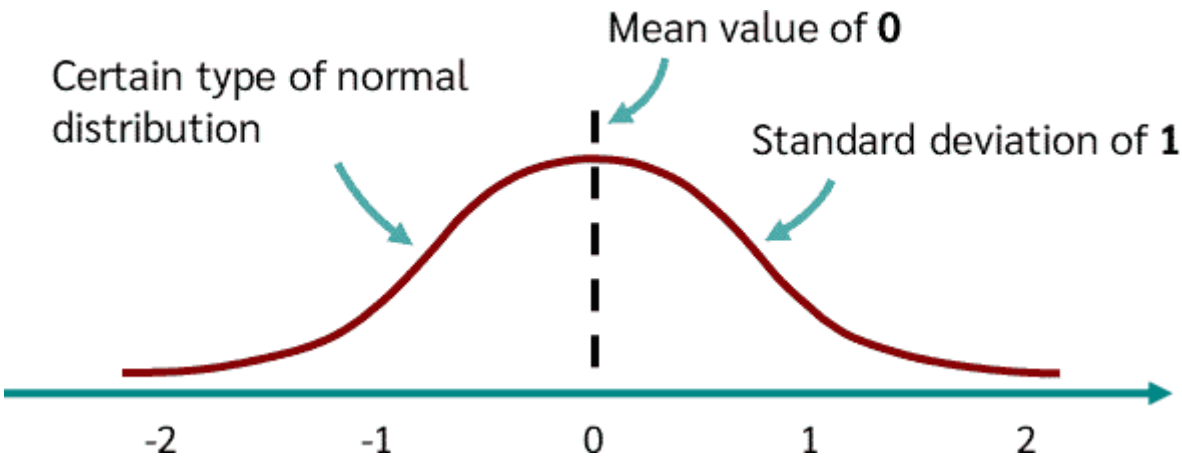
Now, of course, we only have a sample that comes from a specific population. But if the data is normally distributed and the sample size is greater than 30, then we can use the z-value to say what percentage of patients have a blood pressure lower than 110, for example, and what percentage have a blood pressure higher than 110.



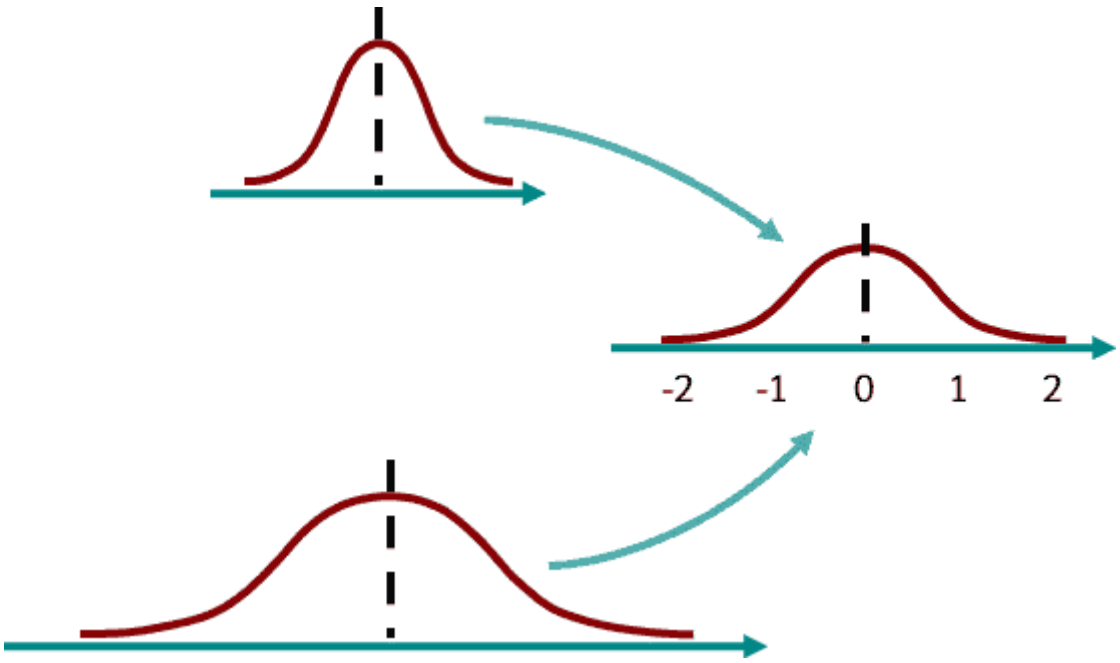
But how does this work? If the initial data is normally distributed, we obtain a so-called standard normal distribution through z-standardization.



The standard normal distribution is a specific type of normal distribution with a mean value of 0 and a standard deviation of 1.



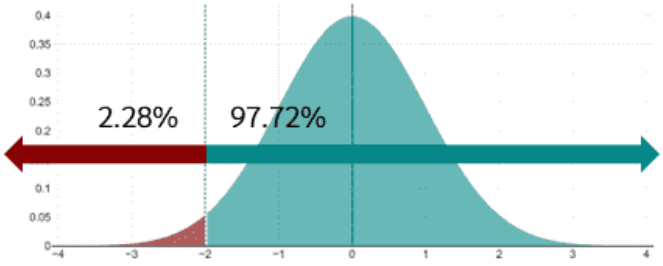
The special feature is that any normal distribution, regardless of its mean or standard deviation, can be converted into a standard normal distribution.



Since we now have a standardized distribution, all we really need is a table that tells us what percentage of the values are below this value for as many z-values as possible. You can find such a table on our website datatab.det (Table of the z-distribution).

Now, of course, the question is how to read this **z-value-table**?

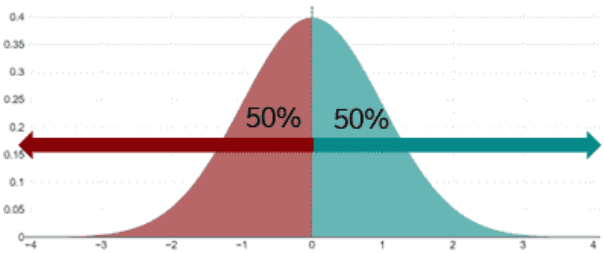
If, for example, we have a z-value of -2, then we can read a value of 0.0228 from this table.



z	0	0.01	0.02	0.03	
-3.4	0.0003	0.0003	0.0003	0.0003	
-3.3	0.0005	0.0005	0.0005	0.0004	
-3.2	0.0007	0.0007	0.0006	0.0006	0.
-3.1	0.001	0.0009	0.0009	0.0009	0.00
-3	0.0013	0.0013	0.0013	0.0012	0.001
-2.9	0.0019	0.0018	0.0018	0.0017	0.00
-2.8	0.0026	0.0025	0.0024	0.0023	0.(
-2.7	0.0035	0.0034	0.0033	0.0032	(
-2.6	0.0047	0.0045	0.0044	0.0043	0.(
-2.5	0.0062	0.006	0.0059	0.0057	0.00
-2.4	0.0082	0.008	0.0078	0.0075	0.(
-2.3	0.0107	0.0104	0.0102	0.0099	(
-2.2	0.0139	0.0136	0.0132	0.0129	
-2.1	0.0179	0.0174	0.017	0.0166	
-2	0.0228	0.0222	0.0217	0.0212	(
-1.9	0.0287	0.0281	0.0274	0.0268	0.(
-1.8	0.0359	0.0351	0.0344	0.0336	0.03
-1.7	0.0446	0.0436	0.0427	0.0418	0.04
-1.6	0.0548	0.0537	0.0526	0.0516	0
-1.5	0.0668	0.0655	0.0643	0.063	(
-1.4	0.0808	0.0793	0.0778	0.0764	0.
-1.3	0.0968	0.0951	0.0934	0.0918	0.(
-1.2	0.1151	0.1131	0.1112	0.1093	0.1(
-1.1	0.1357	0.1335	0.1314	0.1292	0.12
-1	0.1587	0.1562	0.1539	0.1515	0.1
-0.9	0.1841	0.1814	0.1788	0.1762	0.
-0.8	0.2119	0.209	0.2061	0.2033	0
-0.7	0.242	0.2389	0.2358	0.2327	0.2
-0.6	0.2743	0.2709	0.2676	0.2643	0.261
-0.5	0.3085	0.305	0.3015	0.2981	0.29
-0.4	0.3446	0.3409	0.3372	0.3336	0.3
-0.3	0.3821	0.3783	0.3745	0.3707	0.3(
-0.2	0.4207	0.4168	0.4129	0.409	0.405
-0.1	0.4602	0.4562	0.4522	0.4483	0.444
0	0.5	0.496	0.492	0.488	0.4

This means that 2.28% of the values are smaller than a z-value of -2. As the sum is always 100% or 1, 97.72% of the values are greater.

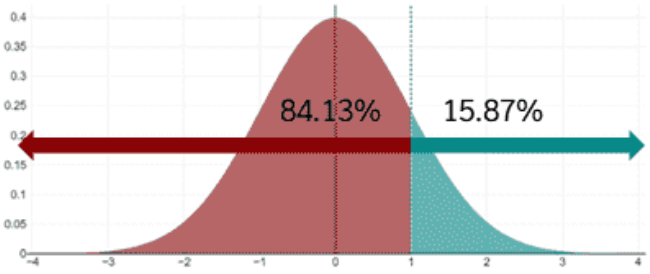
And with a z-value of zero, we are exactly in the middle and get a value of 0.5. Therefore 50% of the values are smaller than a z-value of 0 and 50% of the values are greater than 0. As the normal distribution is symmetrical, we can read off the probabilities for positive z-values exactly.



z-Score of 0

z	0	0.01	0.02	0.03	
-3.4	0.0003	0.0003	0.0003	0.0003	
-3.3	0.0005	0.0005	0.0005	0.0004	
-3.2	0.0007	0.0007	0.0006	0.0006	0.
-3.1	0.001	0.0009	0.0009	0.0009	0.00
-3	0.0013	0.0013	0.0013	0.0012	0.001
-2.9	0.0019	0.0018	0.0018	0.0017	0.00
-2.8	0.0026	0.0025	0.0024	0.0023	0.0
-2.7	0.0035	0.0034	0.0033	0.0032	0
-2.6	0.0047	0.0045	0.0044	0.0043	0.0
-2.5	0.0062	0.006	0.0059	0.0057	0.00
-2.4	0.0082	0.008	0.0078	0.0075	0.0
-2.3	0.0107	0.0104	0.0102	0.0099	0
-2.2	0.0139	0.0136	0.0132	0.0129	
-2.1	0.0179	0.0174	0.017	0.0166	
-2	0.0228	0.0222	0.0217	0.0212	0
-1.9	0.0287	0.0281	0.0274	0.0268	0.0
-1.8	0.0359	0.0351	0.0344	0.0336	0.03
-1.7	0.0446	0.0436	0.0427	0.0418	0.04
-1.6	0.0548	0.0537	0.0526	0.0516	0
-1.5	0.0668	0.0655	0.0643	0.063	0
-1.4	0.0808	0.0793	0.0778	0.0764	0.
-1.3	0.0968	0.0951	0.0934	0.0918	0.0
-1.2	0.1151	0.1131	0.1112	0.1093	0.10
-1.1	0.1357	0.1335	0.1314	0.1292	0.12
-1	0.1587	0.1562	0.1539	0.1515	0.1
-0.9	0.1841	0.1814	0.1788	0.1762	0.
-0.8	0.2119	0.209	0.2061	0.2033	0
-0.7	0.242	0.2389	0.2358	0.2327	0.2
-0.6	0.2743	0.2709	0.2676	0.2643	0.261
-0.5	0.3085	0.305	0.3015	0.2981	0.29
-0.4	0.3446	0.3409	0.3372	0.3336	0.3
-0.3	0.3821	0.3783	0.3745	0.3707	0.3
-0.2	0.4207	0.4168	0.4129	0.409	0.405
-0.1	0.4602	0.4562	0.4522	0.4483	0.444
0	0.5	0.496	0.492	0.488	0.4

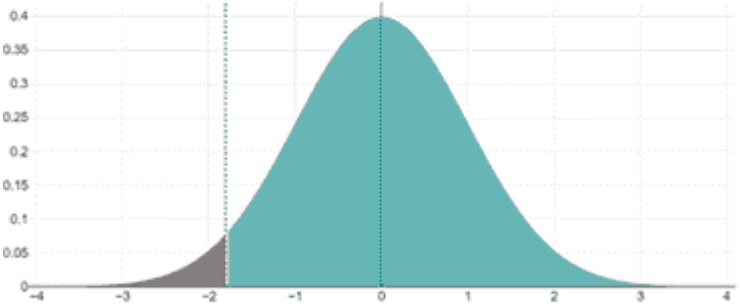
If we have a z-value of 1, we only need to search for -1. However, we must note that in this case we get a value that tells us what percentage of the values are greater than the z-value. So with a z-value of 1, 15.81% of the values are larger and 84.14% of the values are smaller.



z-Score of 1

z	0	0.01	0.02	0.03	
-3.4	0.0003	0.0003	0.0003	0.0003	
-3.3	0.0005	0.0005	0.0005	0.0004	
-3.2	0.0007	0.0007	0.0006	0.0006	0.
-3.1	0.001	0.0009	0.0009	0.0009	0.00
-3	0.0013	0.0013	0.0013	0.0012	0.001
-2.9	0.0019	0.0018	0.0018	0.0017	0.00
-2.8	0.0026	0.0025	0.0024	0.0023	0.0
-2.7	0.0035	0.0034	0.0033	0.0032	0
-2.6	0.0047	0.0045	0.0044	0.0043	0.0
-2.5	0.0062	0.006	0.0059	0.0057	0.00
-2.4	0.0082	0.008	0.0078	0.0075	0.0
-2.3	0.0107	0.0104	0.0102	0.0099	0
-2.2	0.0139	0.0136	0.0132	0.0129	
-2.1	0.0179	0.0174	0.017	0.0166	
-2	0.0228	0.0222	0.0217	0.0212	0
-1.9	0.0287	0.0281	0.0274	0.0268	0.0
-1.8	0.0359	0.0351	0.0344	0.0336	0.03
-1.7	0.0446	0.0436	0.0427	0.0418	0.04
-1.6	0.0548	0.0537	0.0526	0.0516	0
-1.5	0.0668	0.0655	0.0643	0.063	0
-1.4	0.0808	0.0793	0.0778	0.0764	0.
-1.3	0.0968	0.0951	0.0934	0.0918	0.0
-1.2	0.1151	0.1131	0.1112	0.1093	0.10
-1.1	0.1357	0.1335	0.1314	0.1292	0.12
-1	0.1587	0.1562	0.1539	0.1515	0.1
-0.9	0.1841	0.1814	0.1788	0.1762	0.1
-0.8	0.2119	0.209	0.2061	0.2033	0
-0.7	0.242	0.2389	0.2358	0.2327	0.2
-0.6	0.2743	0.2709	0.2676	0.2643	0.261
-0.5	0.3085	0.305	0.3015	0.2981	0.29
-0.4	0.3446	0.3409	0.3372	0.3336	0.3
-0.3	0.3821	0.3783	0.3745	0.3707	0.3
-0.2	0.4207	0.4168	0.4129	0.409	0.405
-0.1	0.4602	0.4562	0.4522	0.4483	0.444
0	0.5	0.496	0.492	0.488	0.4

But what if, for example, we want to read a z-value of -1.81 in the table? We need the other columns for this. We can read a z-value of -1.81 at -1.8 and at 0.01.

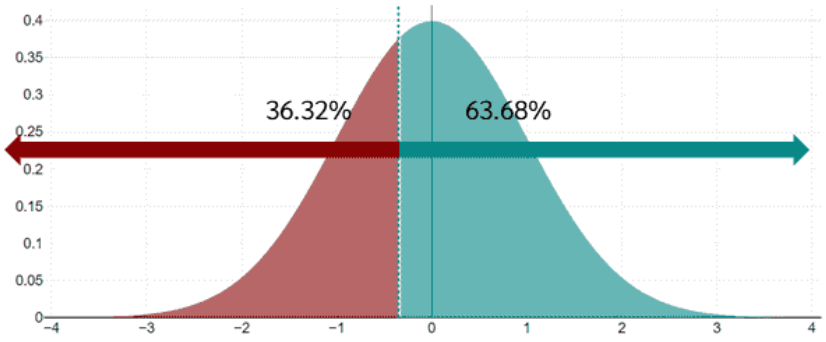


z	0	0.01	0.02	0.03	
-3.4	0.0003	0.0003	0.0003	0.0003	
-3.3	0.0005	0.0005	0.0005	0.0004	
-3.2	0.0007	0.0007	0.0006	0.0006	0.
-3.1	0.001	0.0009	0.0009	0.0009	0.00
-3	0.0013	0.0013	0.0013	0.0012	0.001
-2.9	0.0019	0.0018	0.0018	0.0017	0.00
-2.8	0.0026	0.0025	0.0024	0.0023	0.0
-2.7	0.0035	0.0034	0.0033	0.0032	0
-2.6	0.0047	0.0045	0.0044	0.0043	0.0
-2.5	0.0062	0.006	0.0059	0.0057	0.00
-2.4	0.0082	0.008	0.0078	0.0075	0.0
-2.3	0.0107	0.0104	0.0102	0.0099	0
-2.2	0.0139	0.0136	0.0132	0.0129	
-2.1	0.0179	0.0174	0.017	0.0166	
-2	0.0228	0.0222	0.0217	0.0212	0
-1.9	0.0287	0.0281	0.0274	0.0268	0.0
-1.8	0.0359	0.0351	0.0344	0.0336	0.03
-1.7	0.0446	0.0436	0.0427	0.0418	0.04
-1.6	0.0548	0.0537	0.0526	0.0516	0
-1.5	0.0668	0.0655	0.0643	0.063	0
-1.4	0.0808	0.0793	0.0778	0.0764	0.
-1.3	0.0968	0.0951	0.0934	0.0918	0.0
-1.2	0.1151	0.1131	0.1112	0.1093	0.10
-1.1	0.1357	0.1335	0.1314	0.1292	0.12
-1	0.1587	0.1562	0.1539	0.1515	0.1
-0.9	0.1841	0.1814	0.1788	0.1762	0.1
-0.8	0.2119	0.209	0.2061	0.2033	0
-0.7	0.242	0.2389	0.2358	0.2327	0.2
-0.6	0.2743	0.2709	0.2676	0.2643	0.261
-0.5	0.3085	0.305	0.3015	0.2981	0.29
-0.4	0.3446	0.3409	0.3372	0.3336	0.3
-0.3	0.3821	0.3783	0.3745	0.3707	0.3
-0.2	0.4207	0.4168	0.4129	0.409	0.405
-0.1	0.4602	0.4562	0.4522	0.4483	0.444
0	0.5	0.496	0.492	0.488	0.4

Now let's look at the example about blood pressure again. For example, if we want to know what percentage of patients have a blood pressure below 123, we can use z-standardization to convert a blood pressure of 123 into a z-value, in this case we get a z-value of -0.35.

$$z = \frac{x - \mu}{\sigma} = \frac{123 - 130}{20} = -0.35$$

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-2	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.025	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.063	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.102	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.123	0.121	0.119	0.117
-1	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.166	0.1635	0.1611
-0.8	0.2119	0.209	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.242	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.305	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.281	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.33	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.352	0.3483
-0.2	0.4207	0.4168	0.4129	0.409	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0	0.5	0.496	0.492	0.488	0.484	0.4801	0.4761	0.4721	0.4681	0.4641

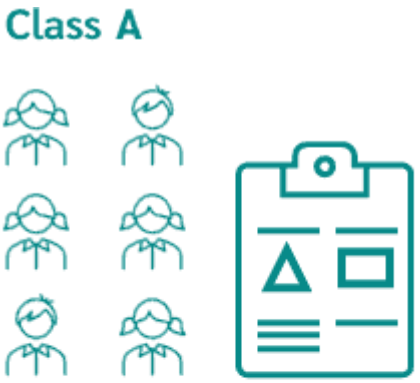


Now we can take the table with the z-distributions and search for a z-value of -0.35. Here we have a value of 0.3632. This means that 36.32 percent of the values are smaller than a z-value of -0.35 and 63.68 percent are larger.

26.4 Compare different data sets with the z-score

However, there is another important application for z-standardization. The z-standardization can help to make values measured in different ways comparable. Here is an example.

Suppose we have two classes, class A and class B, who have written a different test in mathematics.



Mean value

$$z = \frac{x - 70}{10}$$

Standard deviation

Mean value

$$z = \frac{x - 140}{20}$$

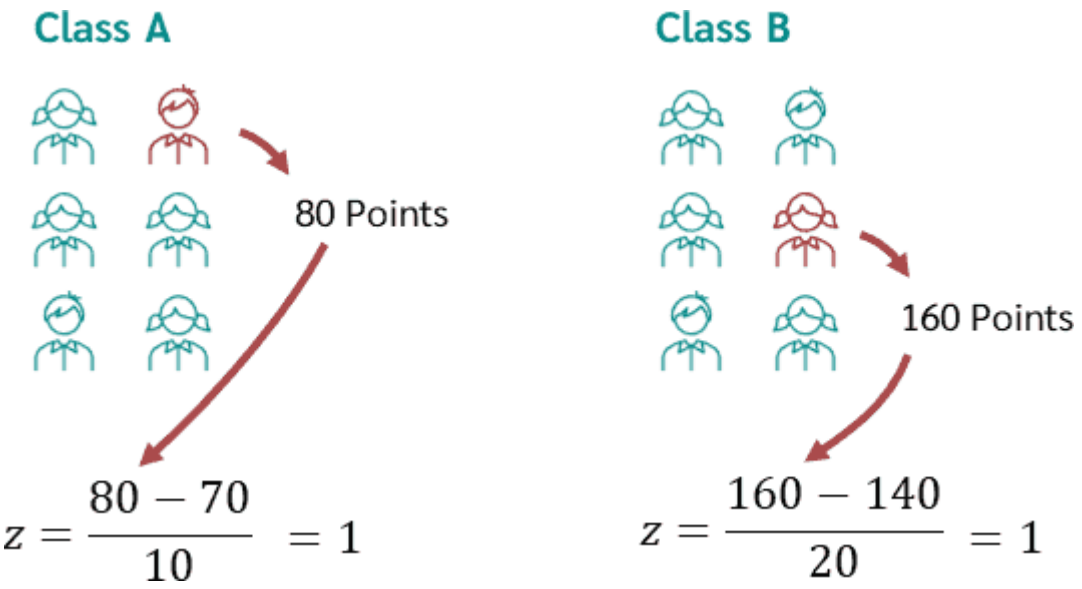
Standard deviation

The tests are designed differently, have a different level of difficulty and a different maximum score.

In order to be able to compare the performance of the pupils in the two classes fairly, we can apply the z-standardization.

The average score or mean score for class A was 70 points with a standard deviation of 10 points. The average score for the test in class B was 140 points with a standard deviation of 20 points.

We now want to compare the performance of Max from class A, who scored 80 points, with the performance of Emma from class B, who scored 160 points.



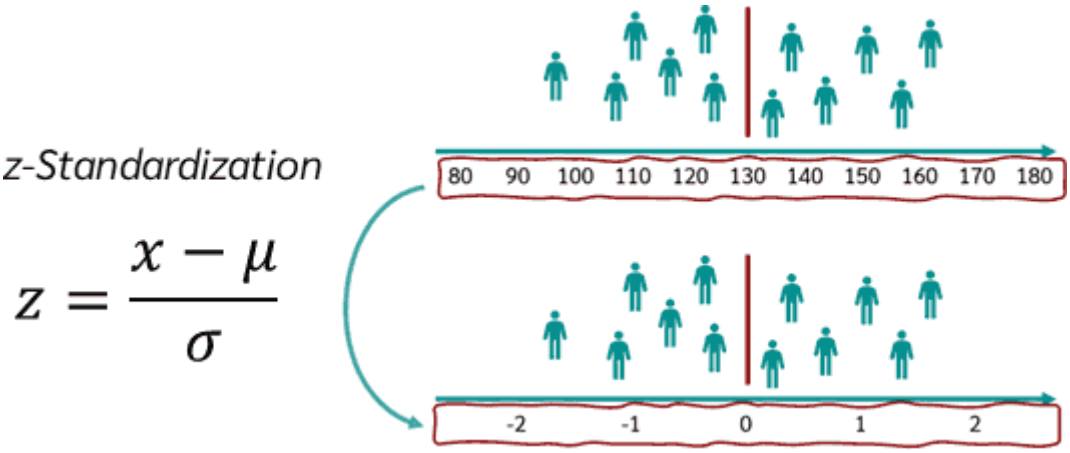
To do this, we simply calculate the z-value of Max and Emma. We enter 80 once for x and get a z-value of 1. Then we enter 160 for x and also get a z-value of 1.

The z-values of Max and Emma are therefore the same. This means that both students performed equally well in terms of average performance and dispersion in their respective classes. Both are exactly one standard deviation above the mean of their class.

26.5 Assumptions z-standardization

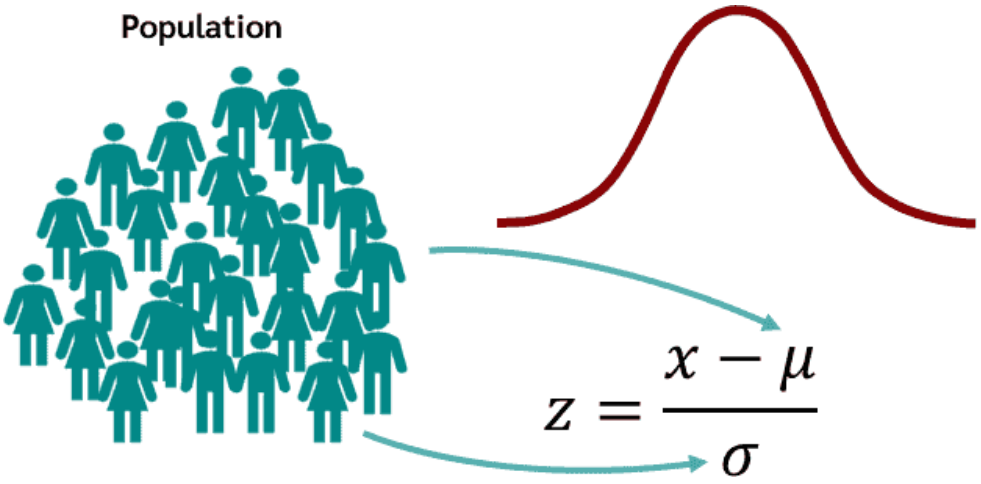
But what about the assumptions? Can we simply calculate a z-standardization and use the table of the standard normal distribution?

The z-standardization itself, i.e. the conversion of the data points into z-values using this formula, is essentially not subject to any strict conditions. It can be carried out independently of the data distribution.



However, if we use the resulting z-values in the context of the standard normal distribution for statistical analyses (e.g. for hypothesis tests or confidence intervals), certain assumptions must be met.

The z-distribution assumes that the underlying population is normally distributed and that the mean (μ) and standard deviation (σ) of the population are known.



However, as you never have the entire population in practice and the mean value and standard deviation are usually not known, this requirement is of course often not met. Fortunately, however, there is an alternative assumption.

Although the z-distribution is defined for normally distributed populations, the central limit theorem can be applied to large samples. This theorem states that the distribution of the sample approaches a normal distribution if the sample size is greater than 30. Therefore, if the sample is larger than 30, the standard normal distribution can be used as an approximation and the mean and standard deviation can be estimated using the sample.

When the standard deviation is estimated from the sample, s is usually written instead of σ and x dash instead of mu for the mean.

$$z = \frac{x - \mu}{\sigma} \qquad z = \frac{x - \bar{x}}{s}$$

References

Backhaus, Klaus, Erichson, Bernd and Plinke, Wulff (2015): *Multivariate analysis methods: an application-oriented introduction.* Berlin/Heidelberg: Springer.

Backhaus, Klaus, Erichson, Bernd, Advanced multivariate analysis methods.

Bortz, Jürgen and Nicola Döring (2015): *Research methods and evaluation in the social and human sciences.* Berlin/Heidelberg: Springer.

Bortz, Jürgen and Christoph Schuster (2010): *Statistics for human and social scientists*. 7th ed. Heidelberg: Springer.

Bühl, Achim. 2012. SPSS 20. *introduction to modern data analysis.* Munich: Pearson Verlag.

Fahrmeier, Ludwig et al. (2016): *Statistics: the road to data analysis*. Heidelberg: Springer.

Diaz-Bone, Rainer (2018): *Statistics for sociologists*. UTB basics. Konstanz: UVK.

Dieckmann, Andreas. *Empirical social research: foundations, methods, applications.* Hamburg: Rowohlt Verlag.

Fromm, Sabine. 2012. *data analysis with SPSS for advanced students 2: multivariate procedures for cross-sectional data.* Wiesbaden: VS Verlag.

Grabinger, Benno (2018): *Fit fürs Studium - Statistik: All basics explained in an understandable way. Suitable for courses of study with statistical methods: VWL, BWL, computer science, etc.* Bonn: Rheinwerk Verlag.

Häder, Michael (2010): *Empirical social research. An introduction*. Wiesbaden: VS Verlag.

Kuckartz, Udo et al. (2013): *Statistics. An understandable introduction*. Wiesbaden: Springer VS.

White, Christel (2019): *Basic knowledge of medical statistics.* Heidelberg: Springer.